# Chapter 1

# Introduction aux séances de TP de biostatistique

Nous allons, au travers de ces séances de travaux pratiques, travailler essentiellement avec 2 logiciels: le premier, Excel, est un logiciel créé par Microsoft, et à ce titre est un logiciel pour lequel une licence est requise. Signalons toutefois que des alternatives existent, certaines étant gratuites (Open Office, par exemple). L'usage de ces autres logiciels est assez souvent similaire à l'utilisation d'Excel, avec toutefois quelques différences notables. Cette famille de logiciels est la famille des "tableurs". Leur fonction majeure est de permettre de faire des calculs, y compris statistiques, ce qui nous intéresse évidemment ici. Ils permettent aussi de construire des graphiques de manière simple, une autre fonctionnalité qui est d'intérêt pour notre cours. Le second logiciel, nommé R, est également un logiciel de calcul avec des fonctionnalités statistiques et graphiques. Il comporte certaines fonctionnalités similaires à celles d'Excel, et le choix de l'un ou l'autre pour résoudre certains exercices sera une question de choix personnel. Moins intuitif, il a l'avantage d'être parfois plus simple qu'Excel (même si ce point est un sujet de discussion avec les étudiants...) et il est gratuit... Dans les exercices résolus qui suivent, nous proposerons souvent une résolution avec chaque logiciel, afin que vous puissiez vous faire votre propre idée sur les avantages et inconvénients de chaque approche.

## 1.1 Introduction Excel (Séance 1)

De nombreux didacticiels existent sur internet, comme par exemple `http://www.Excel-online.net/index2.htm`. Il est évidemment conseillé de parcourir ce type de sites pour se familiariser avec le logiciel, les quelques explications qui suivent n'étant qu'un bref aperçu des possibilités offertes.

### 1.1.1 Cellules

Un document Excel est caractérisé par une grille dans laquelle les utilisateurs vont introduire les données et les informations qu'ils veulent manipuler. Les

cases constituant cette grille sont appelées *cellules*. Chaque cellule est caractérisée par une série de propriétés, parmi lesquelles les plus importantes sont:

- les coordonnées de la cellule, souvent écrites sous la forme d'une référence de ligne et une référence de colonne. Dans le format le plus courant, les colonnes sont désignées par des lettres ("A", "B", $\cdots$, "Z", "AA", "AB", $\cdots$) et les lignes par des nombres (1, 2, $\cdots$). Pour désigner une cellule, on parlera par exemple de la "cellule C4" pour désigner la cellule à l'intersection de la troisième colonne ("C") et de la quatrième ligne ("4").
- la valeur de la cellule. Il s'agit de la valeur (numérique ou alphabétique). Cette valeur a été introduite par l'utilisateur, soit directement, soit par l'intermédiaire d'une formule (voir plus bas).
- le formatage de la cellule, lui-même constitué de plusieurs caractéristiques:
  - les dimensions (largeur et hauteur),
  - la couleur de fond,
  - la couleur de la police,
  - la police de caractère,
  - la bordure,
  - et beaucoup d'autres...
- la formule de la cellule. Les formules Excel forment en fait le cœur du logiciel. Une formule est une expression faisant intervenir différents éléments tels que des nombres, d'autres cellules, des fonctions d'Excel, etc... L'évaluation de l'expression fournit une valeur qui deviendra la valeur de la cellule. On donnera de nombreux exemples de formules dans la suite.

La manipulation des propriétés d'une cellule se fait de manière assez simple:

- l'ajout (ou la modification ou la suppression) d'une valeur se fait en cliquant sur une cellule et en tapant la valeur au clavier,
- les manipulations sur le format de la cellule peuvent se faire via des menus généraux ou locaux (obtenus en cliquant sur la cellule puis en faisant un clic droit), comme montré dans la figure 1.1. Signalons dès à présent que certaines fonctionnalités et que l'apparence des menus et des feuilles peuvent légèrement différer entre les versions d'Excel.

- l'ajout (ou la modification ou la suppression) d'une formule se fait en sélectionnant la cellule visée (en cliquant dessus), puis en tapant la formule, qui apparaitra dans la cellule (pendant l'écriture de celle-ci. Une fois validée, elle sera remplacée dans la cellule par la valeur) et dans la "barre de formule" (voir la figure 1.2). Les formules commencent par le symbole "=", suivi d'une expression Excel. Dans la figure 1.2, la formule "$=10*5+3$" effectue une simple opération arithmétique, dont le résultat (53) deviendra la valeur de la cellule une fois la formule validée (en tapant sur la touche <ENTER>). Les formules sont examinées dans la la section qui suit.

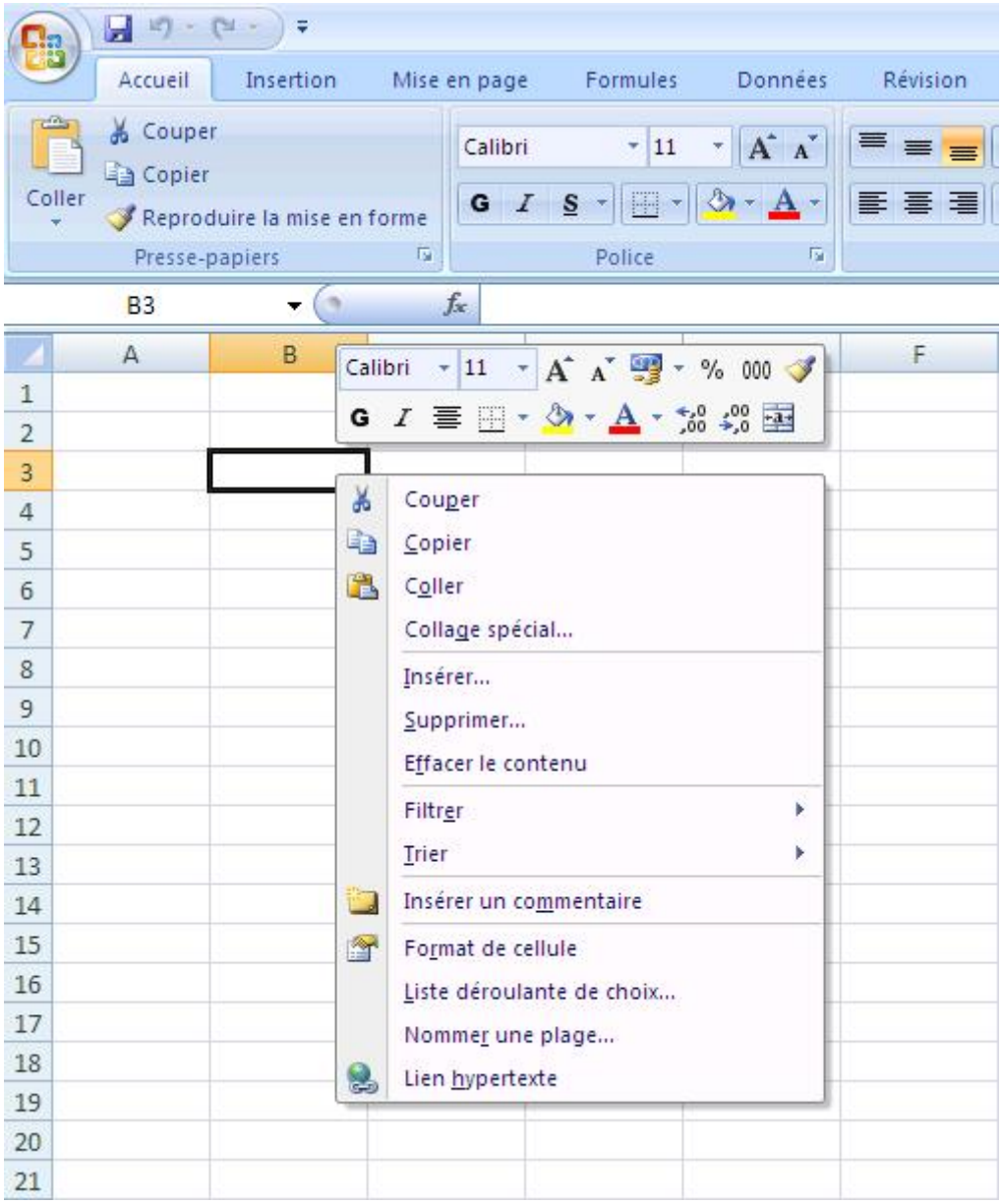



Figure 1.1: Un exemple de menu local. La petite fenêtre du haut permet de préciser la police, la taille des caractères, le centrage dans la cellule, les couleurs de fond et du texte, la bordure, la mise en gras ou en italique, etc... D'autres options sont également disponibles dans la fenêtre inférieure, via l'option "Format de cellule".

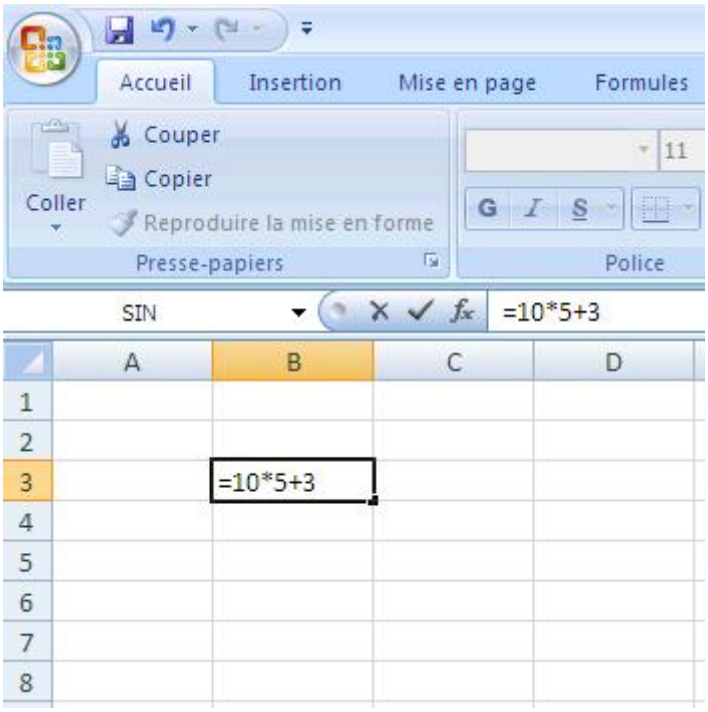



Figure 1.2: Un exemple de formule. Le texte de la formule apparait simultanément dans la cellule visée et dans la "barre de formule". Une fois validée, la formule reste visible dans la barre de formule, mais est remplacée par la valeur correspondante au niveau de la cellule.

### 1.1.2 Plages

Une plage est un ensemble de cellules. Il est parfois utile de sélectionner une plage entière pour faire une manipulation (par exemple, pour changer la couleur de fond de toutes les cellules de la plage d'un seul coup. Pour sélectionner une plage, il existe plusieurs possibilités:

- si la plage à sélectionner est rectangulaire, on peut cliquer sur la cellule occupant le coin supérieur gauche de la cellule (ce qui sélectionne cette cellule), relâcher le clic, déplacer le pointeur de souris au dessus de la cellule inférieure droite de la plage, appuyer sur majuscule et cliquer. La plage sélectionnée apparait alors en gris (voir la figure 1.3). Alternativement, après avoir cliqué sur la cellule du coin supérieur gauche, on peut déplacer le curseur en maintenant le clic (gauche), puis relâcher une fois le coin inférieur droit de la plage atteint.

- une alternative est de sélectionner les cellules constituant la plage de manière individuelle en cliquant sur la première, puis en déplaçant le pointeur de la souris sur la cellule suivante à sélectionner, en appuyant sur la touche <Ctrl> puis, tout en gardant la touche enfoncée, en cliquant. On répète ensuite cette manipulation pour chacune des autres cellules qu'on souhaite inclure (voir la figure 1.4).

- il est également possible de combiner les deux approches.

### 1.1.3 Formules

Une formule est donc une propriété d'une cellule qui permet d'en calculer la valeur. Un des intérêts majeurs d'Excel provient de l'utilisation des formules. En effet, l'utilisation d'une formule dans une cellule permet de passer d'un contenu statique (lorsqu'on entre simplement une valeur dans la formule, cette valeur reste fixe jusqu'à ce que l'utilisateur décide de le modifier manuellement) à un contenu dynamique: si on fait dépendre, via une formule, le contenu d'une cellule A du contenu d'une autre cellule B, toute modification apportée à la valeur de cette cellule B conduira à une modification de la valeur de la cellule A. Par exemple, supposons qu'on introduit la formule "=A1*A1" dans la cellule A2. Dans cette formule:

- le symbole "=" indique qu'il s'agit d'une formule (et non d'une chaine de caractères fixe),

- "A1" désigne la valeur de la cellule de coordonnées A1.

La formule signifie donc: "prends le contenu de la cellule A1, multiplie le par le contenu de la cellule A1 et retourne le résultat qui devient la valeur de la cellule A2". En d'autres termes, la cellule A2 a pour valeur le carré de la valeur de la cellule A1. L'exemple qui suit montre une utilisation simple de cette fonctionnalité.

**Exercice résolu 1.1.1**

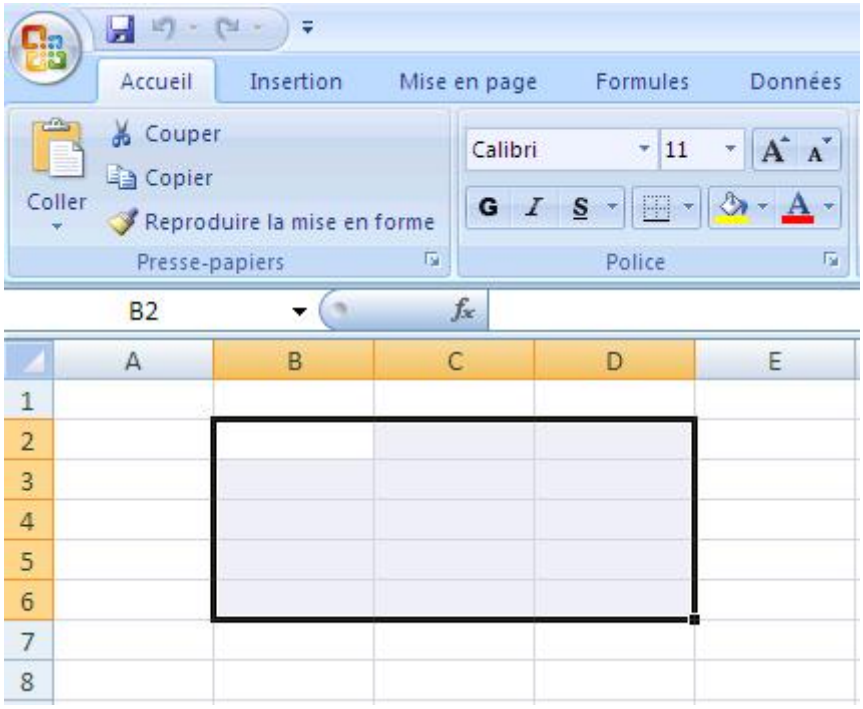



Figure 1.3: Un exemple de sélection de cellules. La plage sélectionnée est rectangulaire et s'étend de B2 à D6, ce qui s'écrit B2:D6 dans le jargon Excel.

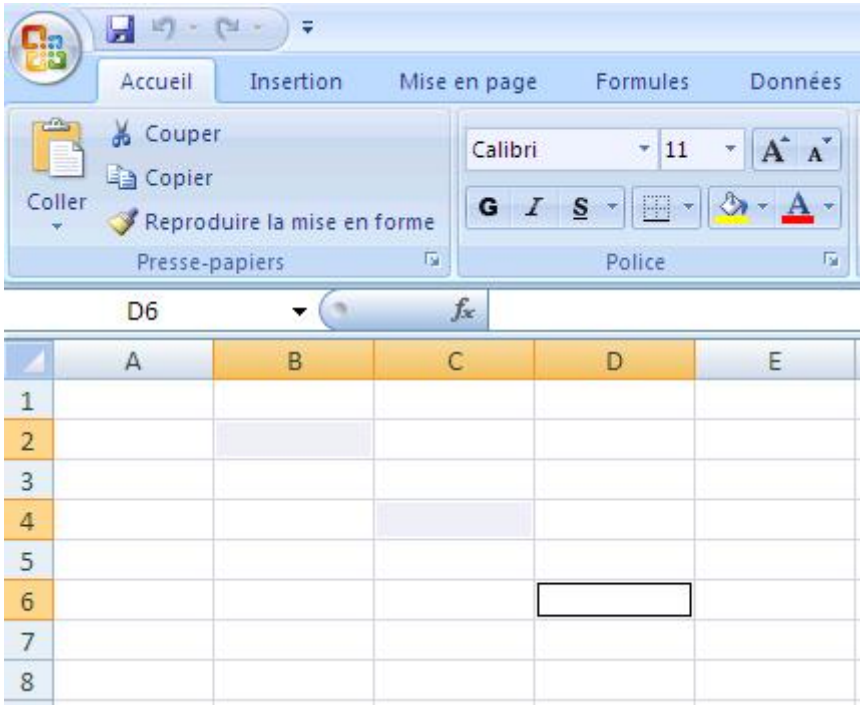



Figure 1.4: Un autre exemple de sélection de cellules. La plage sélectionnée concerne 3 cellules qui ne se touchent pas et s'écrit B2,C4,D6 dans le jargon Excel.

Calculez les racines d'une équation du second degré $ax^2+bx+c=0$ quelles que soient les valeurs de $a$, $b$ et $c$.

►

Un résultat bien connu (j'espère ?) d'algèbre est que les racines se calculent en calculant tout d'abord $\Delta$:

$$\Delta = b^2 - 4 * a * c$$

Ensuite, 3 cas de figure sont possibles:

- si $\Delta$ est négatif, aucune racine n'existe,
- si $\Delta$ est nul, une seule racine existe: $x = -b/(2 * a)$,
- si $\Delta$ est positif, deux racines existent: $x_1 = (-b - \sqrt{\Delta})/(2 * a)$ et $x_2 = (-b + \sqrt{\Delta})/(2 * a)$.

Pour résoudre cet exercice, nous allons donc introduire les chaines de caractères 'a' dans A1, 'b' dans A2, 'c' dans A3, 'Delta' dans A4, 'x1' dans A5 et 'x2' dans A6 (les apostrophes servent à délimiter la chaine dans le texte mais ne doivent pas être tapées dans Excel). Les valeurs respectives pour a, b et c seront introduites dans les cellules B1 à B3. La valeur de $\Delta$ sera calculée dans B4 à partir des valeurs de a, b et c, ce qui signifie que nous allons introduire une formule dans la cellule B4:

$$= B2 * B2 - 4 * B1 * B3$$

ou, en utilisant l'opérateur accent circonflexe d'Excel pour calculer une puissance:

$$= B2\hat{}2 - 4 * B1 * B3$$

Le calcul des racines doit intégrer les conditions sur $\Delta$ énoncées ci-dessus, ce qui nécessite d'introduire les deux premiers exemples de *fonctions Excel*, les fonctions "RACINE et "SI". Une fonction d'Excel s'écrit en écrivant le nom de la fonction, puis, placée entre parenthèses, une liste d'arguments séparés par ";" (dans la version utilisée ici). La fonction est évaluée et la valeur qui résulte de cette évaluation est utilisée dans l'expression qui contient cette fonction. Pour notre problème, on utilisera la fonction "RACINE" qui a un seul argument et qui retourne la valeur de la racine carrée de cet argument, et la fonction "SI", qui a 3 arguments et s'écrit donc:

$$SI(argument_1; argument_2; argument_3)$$

Les arguments sont les suivants:

- $argument_1$ désigne une expression qui est vraie ou qui est fausse (on parle d'expression booléenne ou logique, ou encore de condition),
- $argument_2$ est une expression Excel qui fournira la valeur de la fonction SI si $argument_1$ est vrai,
- $argument_3$ est une expression Excel qui fournira la valeur de la fonction SI si $argument_1$ est faux.

Pour notre problème, on pourra donc faire dépendre les valeurs de $x_1$ et $x_2$ de la valeur de $\Delta$ via cette fonction "SI" de la manière suivante. Dans B5, on introduit la formule:

$$= SI(B4 < 0; "Pas\ de\ racine"; (-B2 - RACINE(B4))/(2 * B1))$$

Si $\Delta < 0$, alors l'expression "B4<0" est vraie et la fonction retourne le second argument, qui est une chaine de caractère "Pas de racine", qui est la réponse correcte. Si par contre $\Delta \geq 0$, la fonction retourne alors le troisième argument, qui calcule la première racine (qui est aussi la seule si $\Delta$ est nul). Il faut ensuite calculer l'éventuelle seconde racine. La formule à introduire en B6 est cette fois:

$$= SI(B4 <= 0; "Pas\ de\ racine"; (-B2 + RACINE(B4))/(2 * B1))$$

Remarquez le changement au niveau de la condition, et évidemment au niveau du calcul de la seconde racine. Une fois ces formules introduites, votre document Excel vous donne les racines (c'est-à-dire les intersections d'une parabole avec l'axe horizontal) pour n'importe quelle équation du second degré $a*x^2+b*x+c = 0$.

◄

**Références absolues et relatives**

Une fonctionnalité importante dans Excel est la possibilité de "copier-coller" une cellule (ou une plage). Pour cela, on sélectionne la cellule (ou la plage) à recopier, on la "copie" (en utilisant le menu "Edition", en utilisant le menu local ou en tapant <Ctrl>-C (taper la touche <Ctrl>, maintenir la touche enfoncée et taper la touche "C")). On sélectionne ensuite la cellule (ou la plage) de destination et on y "colle" ce qui a été copié (en utilisant le menu "Edition", en utilisant le menu local ou en tapant <Ctrl>-V). Si on effectue cette manipulation sur une cellule contenant une formule, il est intéressant de regarder la formule obtenue dans la cellule où l'on a collé la valeur copiée. Pour illustrer ce qui se passe, supposons qu'on veut élever au carré les données placées dans une colonne de la feuille. Pour fixer les idées, on a placé les valeurs de 1 à 10 dans les cellules A1 à A10 et on souhaiterait obtenir les carrés correspondants dans les cellules de B1 à B10. Pour cela, on peut introduire la formule:

$$= A1\hat{}2$$

dans la cellule B1, puis "copier" la cellule B1, sélectionner la plage (rectangulaire) B2:B10 et "coller". Le résultat est affiché dans la figure 1.5 et montre que la manipulation a fonctionné. Si on examine par exemple la formule de la cellule B2, qui est donc une des cellules dans lesquelles on a "copié" le contenu de B1, on s'aperçoit que la formule est:

$$= A2\hat{}2$$

La formule présente dans la case B1 a donc non seulement été recopiée, mais a de plus été adaptée en remplaçant la référence à la cellule A1 par une référence à la cellule A2. Tout se passe donc comme si Excel comprenait les formules avec des références relatives: quand, dans B1, on fait référence à la cellule A1, Excel

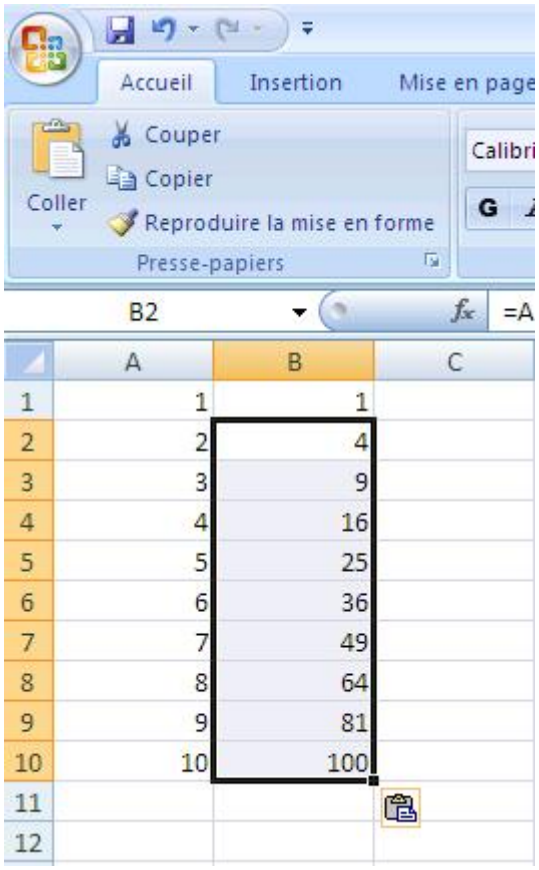



Figure 1.5: Résultat après avoir recopié la formule "= $A1\hat{}2$" de la cellule B1 dans les cellules B2 à B10.

interprète cette référence comme "la cellule directement à la gauche". Quand on recopie la formule de B1 dans B2 (et dans les autres cellules B3:B10), la "cellule directement à gauche" devient la cellule A2, ce qui est bien ce qui est observé. Donc en résumé, quand on recopie une formule dans une cellule située $n$ lignes plus bas dans la feuille Excel, tous les indices de lignes des références aux cellules (comme le "1" de la référence "A1") de la formule sont incrémentés de $n$. Il en va de même si on copie une formule $m$ colonnes à droite de la cellule initiale: les indices de colonne sont augmentés (dans l'ordre alphabétique) de $m$. Ainsi, par exemple, si on recopie la cellule B2 dans la cellule C2:

- l'indice de ligne (2) reste inchangé puisqu'on recopie sur la même ligne,
- l'indice de colonne (A) est augmenté de 1 puisqu'on a recopié 1 colonne plus loin vers la droite: l'indice de colonne devient donc B.

La formule qui apparaitrait dans la cellule C2 serait donc:

$$= B2\hat{}2$$

Si cette utilisation de formules relatives est souvent très utile, il arrive que ce comportement ne soit pas celui qu'on souhaite. Pour illustrer ce fait, supposons qu'on désire calculer la distribution binomiale ayant pour paramètres $n = 5$ (le nombre de répétitions de l'expérience binomiale) et $p = 0.3$ (la probabilité de succès lors de chaque répétition). La distribution binomiale nous permet de calculer la probabilité d'obtenir $r = 0, 1, \cdots, 5$ succès lors des $n = 5$ répétitions. Une fonction d'Excel existe pour effectuer ce type de calculs (cfr plus loin), mais, à titre d'exercice, nous utiliserons ici explicitement la formule binomiale:

$$P(r) = C_n^r * p^r * (1-p)^{n-r}$$

où:

$$C_n^r = \frac{n!}{r! * (n-r)!}$$

Commençons par introduire nos paramètres dans la feuille Excel. Pour que les choses soient claires, nous entrons:

- p dans la cellule A1
- 0,3 dans la cellule B1
- n dans la cellule A2
- 5 dans la cellule B2
- r dans la cellule A3
- P(r) dans la cellule B3

Ensuite, nous allons faire apparaître les valeurs possibles de $r$ (donc, les valeurs de 0 à 5) dans les cellules A4 à A9, et les probabilités correspondantes dans les cellules B4 à B9. Commençons par les valeurs de $r$. Il est évidemment possible d'écrire les chiffres de 0 à 5 dans les cellules A4 à A9, mais une solution plus élégante (et plus efficace, si, au lieu d'avoir $n = 5$, on avait $n = 100$, par exemple...) est de procéder à l'aide de formules. On procède comme suit:

- on introduit la valeur 0 dans la cellule A4,
- on introduit la formule "$= A4 + 1$" (sans les guillemets) dans la cellule A5,
- on "copie" la cellule A5,
- on sélectionne la plage A6:A9,
- on "colle".

A ce stade, vous devriez comprendre le résultat de ce "collage". Remarquez également l'efficacité si $n$ avait été égal à 100: on aurait simplement sélectionné la plage A6:A105 au lieu de la plage A6:A9, et le tour était joué !
Passons à présent au calcul des probabilités. En employant la formule rappelée ci-dessus, et la fonction FACT(n) d'Excel, qui retourne la factorielle de n, on peut facilement calculer la probabilité associée à $r = 0$, puis "copier-coller" la formule pour les autres valeurs de $r$. On commence donc par introduire en B4:

$$= FACT(B2)/(FACT(A4)*FACT(B2-A4))*(B1\hat{}A4)*((1-B1)\hat{}(B2-A4))$$

Le résultat affiché est 0.16807, c'est-à-dire $0.7^5$ (pourquoi ?), ce qui est bien le résultat attendu. Il ne reste donc plus qu'à "copier-coller" la cellule B4 dans les cellules B5:B9 pour compléter le calcul de la distribution. Cette opération conduit à un affichage représenté sur la figure 1.6. Comme on peut le constater sur cette figure, Excel affiche toute une série de messages d'erreurs (sous la forme "$\#VALEUR$!" ou "$\#NOMBRE$!" dans l'exemple) dans les cellules dans lesquelles la cellule B4 a été recopiée. Que se passe-t-il ? On peut comprendre la nature du problème en examinant par exemple la formule de la cellule B5:

$$= FACT(B3)/(FACT(A5)*FACT(B3-A5))*(B2\hat{}A5)*((1-B2)\hat{}(B3-A5))$$

Comme on l'a compris plus haut, toutes les références ont été adaptées suite au recopiage: tous les numéros de lignes ont été augmentés de 1 (puisqu'on a recopié une cellule plus bas la cellule B4), les indices de colonnes restant inchangés (puisqu'on a recopié dans la même colonne). Si c'était bien l'effet escompté pour $r$ (on est passé de la cellule A4 à la cellule A5, faisant passer la valeur de $r$ de 0 à 1), ce n'est évidemment pas ce qui était recherché pour les valeurs de $p$ (on est passé de la cellule B1 à la cellule B2, c'est-à-dire de la valeur 0.3 à la valeur $\cdots$ de 5, qui est en réalité la valeur de $n$) ni pour les valeurs de $n$ (on est passé de la cellule B2 à la cellule B3, c'est-à-dire de la valeur 5 à la valeur $\cdots$ de "P( r )", qui est l'en-tête de la colonne contenant les probabilités à calculer !). On comprend dès lors pourquoi Excel a des difficultés pour effectuer ce calcul, et affiche un message d'erreur. En réalité, ce qu'on aurait souhaité, c'est que la formule prenne la forme:

$$= FACT(B2)/(FACT(A5)*FACT(B2-A5))*(B1\hat{}A5)*((1-B1)\hat{}(B2-A5))$$

En d'autres termes, l'indice de la valeur de $r$ est bien passé de 4 à 5, alors que les indices de $p$ et de $n$ sont restés figés à leur valeur dans la formule initiale, soit B1 et B2, respectivement. Est-il possible de "figer" certains indices dans la formule ? La réponse est oui: il suffit de précéder ces indices du symbole $.

Les coordonnées ainsi annotées deviennent alors absolues, et non plus relatives. Ainsi, si on remplace la formule dans la cellule B4 par la formule suivante:

$$= FACT(B\$2)/(FACT(A4)*FACT(B\$2-A4))*(B\$1\hat{}A4)*((1-B\$1)\hat{}(B\$2-A4))$$

et qu'on "copie-colle" cette formule dans les cellules B5:B9, le résultat est cette fois correct (figure 1.7). L'examen de la formule de la cellule B5 fournit:

$$= FACT(B\$2)/(FACT(A5)*FACT(B\$2-A5))*(B\$1\hat{}A5)*((1-B\$1)\hat{}(B\$2-A5))$$

ce qui montre que les indices de lignes de $p$ et de $n$ on bien été préservés par le placement du symbole \$. Signalons encore que le placement du symbole \$ devant l'indice de ligne ne gèle que la ligne: la colonne reste référencée de manière relative. Si on souhaite geler la colonne, il faut précéder l'indice de colonne du symbole \$ également. Et donc, les adresses $B1$, $B\$1$, $\$B1$ et $\$B\$1$ font toutes référence à la même cellule, mais la recopie de la formule contenant cette référence dans d'autres cellules aura potentiellement un effet différent sur le résultat des calculs, comme l'a illustré l'exemple précédent et comme va encore le démontrer l'exemple qui suit. Signalons enfin que des appuis successifs sur la touche F4 permet de placer les symboles \$ dans la référence en cours.

**Exercice résolu 1.1.2**

Sachant qu'il y a 60% de labradors dorés, 30% de labradors noirs et 10% de labradors chocolat. Donnez la distribution (trinomiale) correspondant aux compositions possibles de groupes de 5 labradors pris au hasard dans la population. Tout d'abord, on peut représenter chaque composition possible par un triplet $(r_D, r_N, r_C)$ où $r_D$, $r_N$ et $r_C$ représentent le nombre de labradors dorés, noirs et chocolats, respectivement. Évidemment, $r_D + r_N + r_C = 5$, ce qui revient à dire que $r_C = 5 - r_D - r_N$. Le triplet peut donc s'écrire $(r_D, r_N, 5 - r_D - r_N)$, qui ne dépend que de $r_D$ et $r_N$. On peut donc représenter les situations possibles par un tableau à doubles entrées, avec le nombre de labradors dorés comme en-têtes de lignes et le nombre de labradors noirs comme en-têtes de colonnes. Cette représentation est illustrée dans la figure 1.8. Il reste à calculer les probabilités associées aux 21 situations possibles. On utilise la formule de la distribution trinomiale, qui s'écrit, dans le cas présent:

$$P(r_D, r_N) = \frac{5!}{r_D! * r_N! * (5 - r_D - r_N)!} * p_D^{r_D} * p_N^{r_N} * (1 - p_D - p_N)^{5 - r_D - r_N}$$

Pour arriver aux probabilités recherchées, on peut introduire la formule suivante dans la cellule C6, puis l'étendre aux autres cellules admissibles du tableau:

$$= fact(5)/(fact(\$B6)*fact(C\$5)*fact(5-\$B6-C\$5))*\$B\$1\wedge\$B6*\$B\$2\wedge C\$5*\$B\$3\wedge(5-\$B6-C\$5)$$

Le résultat de ces manipulations est montré dans la figure 1.9.

**Exercice 1.1.3**

Les 3 races bovines suivantes sont présentes en Belgique: Pie Noire (PN: 50%), Blanc-Bleu Belge (BBB: 30%) et Pie Rouge (PR: 15%) (les 5% restants sont constitués de races diverses). Sachant que des échantillons de 10 individus sont

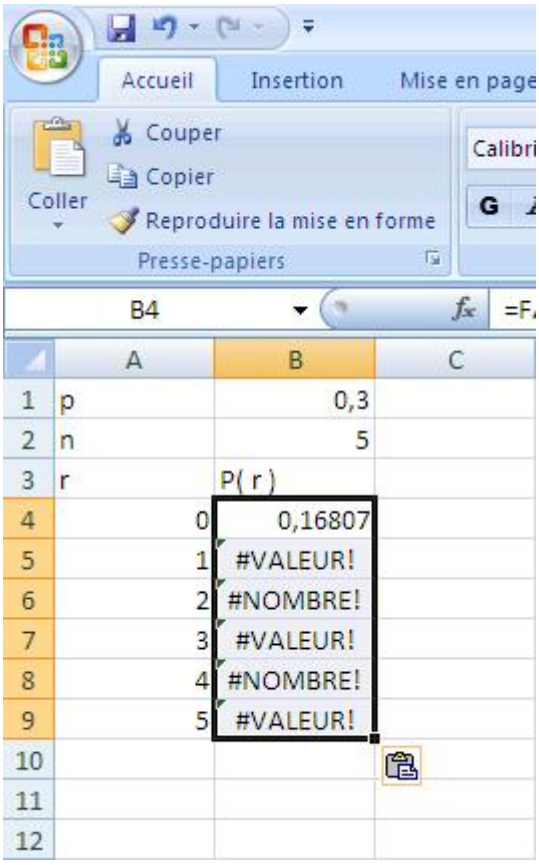



Figure 1.6: Résultat après avoir recopié la formule binomiale de la cellule B4 avec des références relatives dans les cellules B5 à B9.

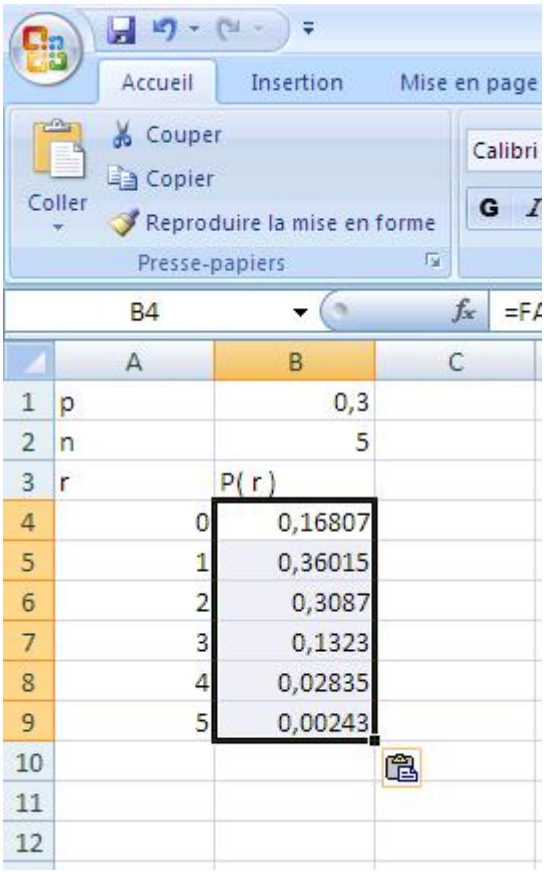



Figure 1.7: Résultat après avoir recopié la formule binomiale avec des références absolues de la cellule B4 dans les cellules B5 à B9.

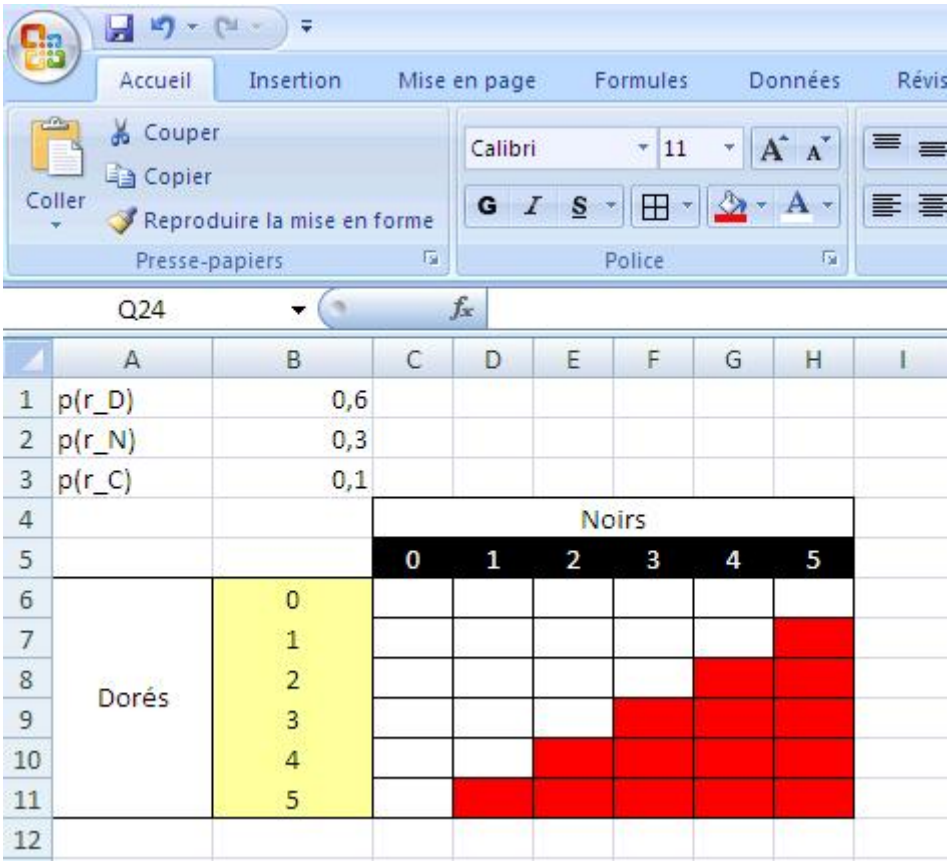
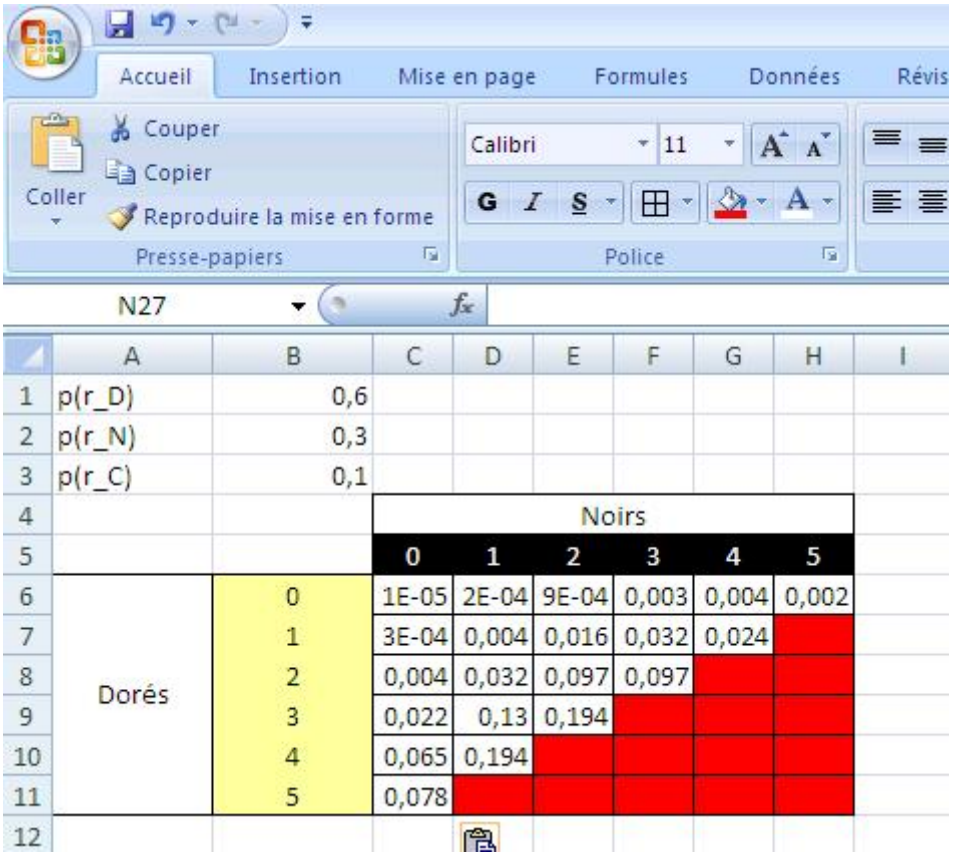



Figure 1.8: Représentation des combinaisons possibles pour les échantillons de 5 labradors. Les cases rouges correspondent à des situations qui ne sont pas possibles.




| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | p(r_D) | 0,6 | | | | | | | |
| 2 | p(r_N) | 0,3 | | | | | | | |
| 3 | p(r_C) | 0,1 | | | | | | | |
| 4 | | | Noirs | | | | | | |
| 5 | | | 0 | 1 | 2 | 3 | 4 | 5 | |
| 6 | Dorés | 0 | 1E-05 | 2E-04 | 9E-04 | 0,003 | 0,004 | 0,002 | |
| 7 | | 1 | 3E-04 | 0,004 | 0,016 | 0,032 | 0,024 | | |
| 8 | | 2 | 0,004 | 0,032 | 0,097 | 0,097 | | | |
| 9 | | 3 | 0,022 | 0,13 | 0,194 | | | | |
| 10 | | 4 | 0,065 | 0,194 | | | | | |
| 11 | | 5 | 0,078 | | | | | | |
| 12 | | | | | | | | | |

Figure 1.9: Probabilités associées aux combinaisons possibles pour les échantillons de 5 labradors. La probabilité la plus élevée est obtenue pour les situations $(r_D, r_N, r_C) = (3, 2, 0)$ et $(r_D, r_N, r_C) = (4, 1, 0)$.

constitués de manière aléatoire à partir des 3 races majeures, quelle est la composition de l'échantillon le plus probable, et quelle est sa probabilité ?
►

**Solution**: la composition la plus probable est (PN, BBB, PR) = (6, 3, 1), avec une probabilité valant 0.0887806.
◄

### 1.1.4 Fonctions

Nous avons déjà vu quelques exemples de fonctions plus haut. Nous en rencontrerons d'autres plus loin relatives aux distributions de probabilités. Dans cette section, nous évoquerons quelques fonctions supplémentaires, utiles dans les calculs statistiques.

- la fonction **ALEA** ne prend aucun argument, et génère des nombres réels au hasard de manière uniforme entre 0 et 1. Signalons toutefois que, bien que la fonction ne prend aucun argument, il est nécessaire de spécifier les parenthèses (dont l'intérieur sera donc vide). Un appel se fait donc par la formule:

  $$= ALEA()$$

  Une variante de cette fonction existe, qui permet de générer de manière aléatoire et uniforme, des nombres entiers dans un intervalle spécifié par une valeur minimale et une valeur maximale. Cette fonction s'écrit

  $$ALEA.ENTRE.BORNES(Min; Max)$$

  où *Min* et *Max* désignent les limites inférieures et supérieure de l'intervalle visé. Par exemple, pour simuler un jet de dés à 6 faces avec Excel, on peut taper la formule:

  $$= ALEA.ENTE.BORNES(1; 6)$$

  qui va donc générer des chiffres 1, 2, 3, 4, 5 ou 6 aléatoirement et de manière uniforme.
  Les fonctions de ce type fournissent un résultat (aléatoire) qualifié de "volatile": à chaque re-calcul de la page (c'est-à-dire, chaque fois que vous modifiez la page ou que vous le demandez explicitement, par exemple en appuyant sur F9), comme la valeur que génère la formule est aléatoire, la valeur de la cellule change.

- la fonction **SOMME** additionne les valeurs (nombres, ou valeurs des cellules) qui lui sont communiquées comme paramètres. Le nombre de paramètres peut aller de 1 à 255, selon les besoins. Chaque paramètre peut être un nombre, une cellule ou une plage de cellules. Les paramètres sont séparés par une séparateur, qui est un ";" dans la version utilisée dans ces notes (comme pour la fonction SI vue plus haut). Ainsi, par exemple, la formule:

  $$= SOMME(A1 : B10; D20 : D30; F1; 20)$$

  additionne les valeurs de toutes les cellules de la plage rectangulaire allant de A1 à B20, puis celles allant de D20 à D30, ajoute la valeur de la cellule

F1, puis ajoute 20 au résultat et le retourne. Signalons que les cellules vides ou non numériques ne sont pas considérées dans le calcul et peuvent donc figurer dans les plages qu'on additionne sans perturber les calculs.

- la fonction **PRODUIT** multiplie ses paramètres.
- la fonction **SOMME.CARRES** additionne les carrés de ses paramètres.
- la fonction **SOMMEPROD** prend pour arguments des plages de dimensions identiques, effectue le produit des éléments correspondants des différentes plages et additionne les produits ainsi obtenus.
- les fonctions **MOYENNE**, **MEDIANE**, **VAR**, **ECARTYPE**, **MAX**, **MIN** fonctionnent comme la fonction SOMME en effectuant le calcul qu'indique clairement leur nom.
- la fonction **CENTILE** prend 2 arguments: une plage de données et une valeur comprise entre 0 et 1 indiquant le centile qui est désiré. Ainsi, par exemple, pour obtenir le percentile 90 dans un échantillon de 1000 valeurs présentes dans les cellules A1 à A1000, on peut taper la formule:

$$= CENTILE(A1 : A1000; 0,9)$$

**Exercice 1.1.4**

Nous allons commencer par générer un lot de 100 données aléatoires, qui pourraient simuler des mesures de tailles faites sur un échantillons de 100 chiens d'une race donnée. La procédure donnée ici sera explicitée plus tard dans les séances de travaux pratiques.

- Dans la case A1 de la feuille, tapez la formule suivante:

$$= ARRONDI(LOI.NORMALE.INVERSE(ALEA(); 50; 10); 1)$$

- Copiez cette cellule,
- Sélectionnez la plage A1:J10 (soit, 100 cellules),
- Collez,
- Copiez la plage,
- Faites un clic droit sur la plage, sélectionnez "Collage spécial" et choisissez "Valeurs" dans la fenêtre qui apparait.

Les deux dernières instructions ont pour objet de remplacer la formule de la cellule par la valeur aléatoire générée. Comme, lors d'un éventuel re-calcul de la page, il n'y a plus de formule à évaluer, les valeurs aléatoires générées restent fixes: on peut donc faire les calculs sans que les résultats changent à chaque manipulation. Pour ces 100 données, on demande de calculer la moyenne, la variance, l'écart-type, l'étendue, les quartiles et le troisième décile.

▶

**Solution**: comme les données sont aléatoires, on ne peut fournir ici qu'un exemple de solution. Vos résultats devraient être proches de ceux fournis ici à titre d'exemple:

- Moyenne: 51.479
- Variance: 126.926
- Écart-type: 11.266
- Étendue: 46.6
- Quartiles: $(Q1, Q2, Q3) = (43.925, 51.300, 59.125)$
- Troisième décile: 45.170

Signalons que les valeurs théoriques (obtenues à partir de la distribution normale qui a servi à simuler les données sont respectivement 50, 100, 10, $+\infty$, 43.255, 50, 56.745 et 44.756.
◀

### 1.1.5 Les macros (Paragraphe facultatif, donné pour information)

Au delà de toutes les fonctionnalités déjà évoquées, Excel dispose également de possibilités de programmation qui augmentent encore sa puissance. Pour ce qui nous concernera ici, les possibilités de programmation nous serviront essentiellement à faire répéter un grand nombre de fois (par exemple, 1000 fois) des calculs en changeant les données à chaque fois afin d'obtenir non pas une, mais bien un grand nombre de solutions, afin d'être en mesure d'étudier la distribution de ces solutions. Ces programmes qu'on peut faire exécuter par le logiciel s'appellent des macros. Nous allons commencer par accéder aux macros dans Excel, puis illustrer l'utilisation des macros sur un exemple simple.

#### Editeur de macros

Pour écrire une macro, une procédure simple consiste à utiliser l'option "Macros" du menu "Affichage". En cliquant sur le $\bigtriangledown$, le petit menu qui s'affiche permet de choisir l'option "Enregistrer une macro" (voir la figure 1.10). Si on choisit l'option "Enregistrer une macro", Excel demande un nom pour la macro à créer, nom qui est quelconque (à quelques contraintes près). Par exemple, nous pourrions donner le nom "Couleurs" (pour une raison qui va apparaître tout de suite). Une fois le nom validé (en tapant sur le bouton "Ok"), Excel commence à enregistrer toutes les commandes que nous exécutons dans Excel dans la macro "Couleurs". A titre d'exemple, nous allons sélectionner la case A1, puis en colorer le fond (avec une couleur de votre choix) avec l'outil "seau de peinture" du menu "Accueil". Une fois ces manipulations effectuées, on peut retourner au menu "Macros" du menu "Affichage", et cette fois sélectionner l'option "Arrêter l'enregistrement". Excel arrête alors d'enregistrer les commandes que nous effectuons dans la macro "Couleurs". On peut voir le contenu de la macro en retournant dans le menu "Macros" et en cliquant cette fois sur "Afficher les macros". Excel propose alors les macros accessibles (sauf si un autre fichier Excel contenant des macros est ouvert, seule la macro "Couleurs" devrait être visible). On sélectionne la macro "Couleurs" et on clique sur "Modifier", ce qui ouvre l'éditeur de macro, qui présente le texte constituant la macro. Le texte devrait ressembler à ceci:

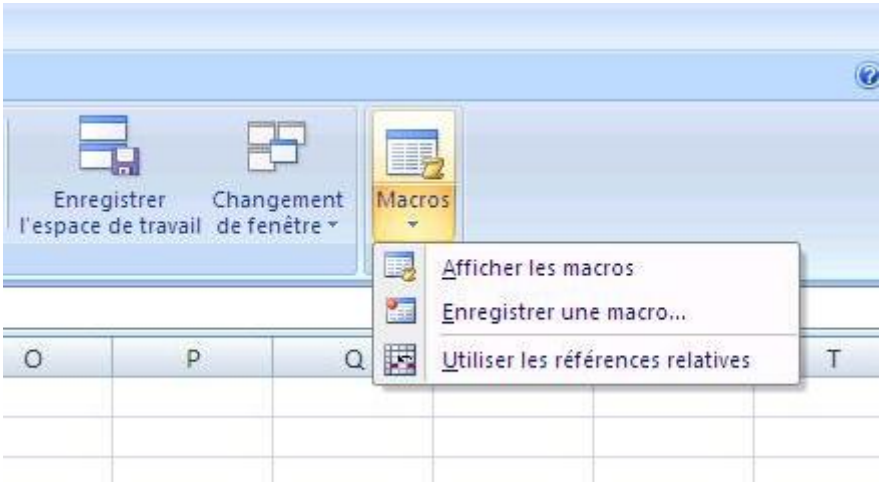



Figure 1.10: Affichage du menu permettant d'enregistrer une macro et d'accéder aux macros éventuellement déjà enregistrées.

```
Sub Couleurs()
'
' Couleurs Macro
'
  Range("A1").Select
  With Selection.Interior
       .Pattern = xlSolid
       .PatternColorIndex = xlAutomatic
       .color = 65535
       .TintAndShade = 0
       .PatternTintAndShade = 0
  End With
End Sub
```

L'interprétation de ce programme est la suivante:

- Les macros commencent par l'instruction:

  $$Sub < nom > (< parametres >)$$

  où $< nom >$ est le nom donné à la macro (Couleurs, dans notre exemple) et $< parametres >$ est une liste de paramètres transmise à la macro (cette liste est vide dans notre exemple: la macro Couleurs ne prend aucun paramètre).

- Les macros se terminent par l'instruction:

  $$End\ Sub$$

  qui terminent le code constituant la macro.

- les lignes qui commencent par une apostrophe sont des commentaires, qui ne seront pas interprétés par Excel. Les commentaires permettent à l'auteur de la macro d'insérer des informations dans le texte de la macro,

s'il le désire. Dans l'exemple, la première et la troisième ligne de commentaire sont vides et sont là pour augmenter la lisibilité du texte. La seconde ligne précise que la macro s'appelle "Couleurs", ce qui était déjà connu. Ces lignes sont facultatives mais sont parfois utiles dans des programmes plus importants pour expliciter le code.

- L'instruction:

$$Range(\text{"}A1\text{"}).Select$$

correspond à l'action menée plus haut qui consistait à sélectionner la case A1. Remarquez la manière de préciser comment on effectue une action (sélectionner) sur un objet (la case "A1"). Pour spécifier l'objet que représente la case A1, on peut donc dire:

$$Range(\text{"}A1\text{"})$$

. Alternativement, on peut spécifier une cellule avec:

$$Cells(< ligne >, < colonne >)$$

Pour spécifier A1, on taperait donc:

$$Cells(1, 1)$$

Une fois l'objet spécifié, on peut faire différentes actions sur cet objet. Dans nos manipulations, nous avons commencé par sélectionner A1. La macro reproduit cette manipulation en disant de considérer l'objet A1 puis de le sélectionner:

$$Range(\text{"}A1\text{"}).Select$$

Une fois l'objet sélectionné, nous avons changé le fond de transparent (couleur 0, par défaut) à jaune (couleur 65535). Les autres propriétés de la cellule sont restées inchangées (c'est-à-dire, avec leur valeur par défaut). Le fond d'une cellule est un objet appelé "Interior" dans le langage des macros. Si on parle du fond de la plage (cellule, ici) sélectionnée, on dira "Selection.Interior", et si on parle de la couleur du fond de la plage sélectionnée, on écrira "Selection.Interior.Color". Cette propriété de la cellule peut être modifiée en lui donnant une nouvelle valeur (par défaut, c'est 0). Il suffit de taper:

$$Selection.Interior.Color = 65535$$

qui change donc cette propriété de la cellule. L'effet visuel de ce changement est de changer la couleur de fond de la sélection de transparent en jaune. Les 4 autres propriétés ("Pattern", "PatterColorIndex", "TintAndShade", "PatternTintAndShade") ne sont pas modifiées (on aurait pu ne pas les spécifier, mais le générateur automatique les écrit en leur réattribuant leur valeur par défaut ("xlSolid", "xlAutomatic", 0 et 0, respectivement).
Pour éviter de taper plusieurs fois "Selection.Interior", comme dans le code équivalent suivant:

```
        Selection.Interior.Pattern = xlSolid
        Selection.Interior.PatternColorIndex = xlAutomatic
        Selection.Interior.color = 65535
        Selection.Interior.TintAndShade = 0
        Selection.Interior.PatternTintAndShade = 0
```

Excel met "Selection.Interior" en évidence dans une instruction commençant par:

$$With\ Selection.Interior$$

et se terminant par:

$$End\ With$$

Toutes les éléments du code commençant par un "." entre ces deux instructions se verront préfixés de "Selection.Interior".

- il existe potentiellement plus de 16 millions de couleurs disponibles, chaque couleur ayant un code. Pour obtenir une couleur déterminée, le plus simple est sans doute:
  - soit de passer par la fonction RGB(R,G,B) qui permet de spécifier l'intensité des composantes rouge (R = Red), verte (G = Green) et bleue (B = Blue), chacune des intensités étant comprise entre 0 et 255. Notez que le jaune de la cellule correspond à une pleine intensité de rouge et de vert: RGB(255,255,0) = 65535.
  - soit de passer par la propriété ColorIndex, qui permet de présenter une palette simplifiée de 56 couleurs. Par exemple, pour obtenir le jaune que nous utilisons pour cet exemple, on pourrait taper:

$$Selection.Interior.ColorIndex = 6$$

### Code simple

Au vu des explications données au paragraphe précédent, il est possible de simplifier le code qui a été généré de manière automatique. Nous allons, pour illustrer cela, colorer les cellules de la plage A1 à A3 en employant les 3 méthodes proposées et en supprimant tout ce qui n'est pas nécessaire:

- la sélection de la cellule pour la manipuler, opération qui n'est pas requise dans les macros (mais qui est incontournable en "manuel"...)
- l'accès aux propriétés non modifiées, qui sont à leur valeur par défaut et n'ont pas besoin d'être modifiées.

Le code pourrait donc s'écrire:

```
Sub Couleurs()
 Cells(1,1).Interior.Color = 65535
 Cells(2,1).Interior.Color = RGB(255, 255, 0)
 Cells(3,1).Interior.ColorIndex = 6
End Sub
```

### Boucle for

Continuant sur l'exemple précédent, on pourrait vouloir savoir à quoi ressemblent les 56 couleurs de la palette simplifiée. Il suffit évidemment de demander à la macro de les afficher:

```
Sub Couleurs()
 Cells(1,1).Interior.ColorIndex = 1
 Cells(2,1).Interior.ColorIndex = 2
 ...
 Cells(56,1).Interior.ColorIndex = 56
End Sub
```

On voit directement qu'il s'agit d'un travail fastidieux, consistant à écrire 56 fois des lignes très similaires sous la forme:

$$Cells(i, 1).ColorIndex = i$$

où $i$ varie de 1 à 56. Excel permet de simplifier ce travail, en utilisant une "boucle for" de la manière suivante:

```
Sub Couleurs()
 For i = 1 to 56
  Cells(i,1).Interior.ColorIndex = i
 Next i
End Sub
```

Comme il est assez clair (?), l'instruction commençant par "For" fait varier une variable nommée (par exemple) i de 1 à 56, et pour chaque valeur de i, exécute le code entre l'instruction "For" et l'instruction "Next" correspondante. Comme l'instruction concernée, dans notre exemple, fait intervenir la variable i, l'instruction exécutée change à chaque passage dans la boucle, ce qui était l'effet désiré pour obtenir les 56 instructions légèrement différentes décrites plus haut.

**Exercice résolu 1.1.5**

Affichez les 56 couleurs sur les 8 premières lignes et 7 premières colonnes de la page, en affichant également pour chacune le code correspondant.

▶

Si on appelle l le numéro de ligne et c le numéro de colonne pour chaque cellule visée, le code couleur correspondant sera:

$$couleur = (ligne - 1) * 7 + c$$

La macro peut donc s'écrire:

```
Sub Couleurs()
 For l = 1 to 8
  For c = 1 to 7
   couleur = (l - 1) * 7 + c
   Cells(l,c).Interior.ColorIndex = couleur
   Cells(l,c).Value = couleur
  Next c
 Next l
End Sub
```

◀


Figure 1.11: Affichage du résultat obtenu à l'issue de l'utilisation de la macro "Couleurs".

### If then else

Continuant à élaborer sur l'exemple précédent, on peut voir que certains codes sont illisibles, parce qu'écrits en noir sur un fond sombre. Il serait bon, pour les cellules concernées, d'utiliser une police claire (par exemple blanche). Pour illustrer l'approche, nous allons effectuer cette manipulation sur les 2 dernières lignes. L'instruction:

$$If < condition > Then < instructions > EndIf$$

permet facilement de tester une condition, et, uniquement si cette condition est remplie, d'effectuer les instructions entre "Then" et "EndIf". Le code est le suivant:

▶

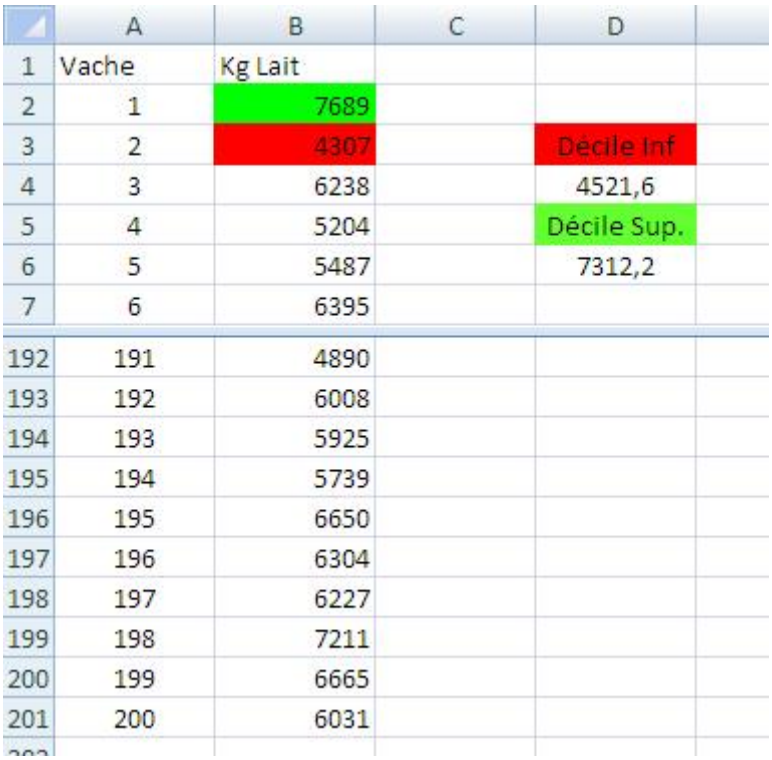
```
Sub Couleurs()
 For l = 1 to 8
  For c = 1 to 7
   couleur = (l - 1) * 7 + c
   Cells(l,c).Interior.ColorIndex = couleur
   Cells(l,c).Value = couleur
   If (l>6) Then
    Cells(l,c).Font.ColorIndex = 2
   End If
  Next c
 Next l
End Sub
```

Le résultat obtenu après l'exécution de cette macro est montré sur la figure 1.11.

◀

## 1.1.6 Exercices

**Exercice résolu 1.1.6**



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Vache | Kg Lait | | |
| 2 | 1 | 7689 | | |
| 3 | 2 | 4307 | | Décile Inf |
| 4 | 3 | 6238 | | 4521,6 |
| 5 | 4 | 5204 | | Décile Sup. |
| 6 | 5 | 5487 | | 7312,2 |
| 7 | 6 | 6395 | | |
| 192 | 191 | 4890 | | |
| 193 | 192 | 6008 | | |
| 194 | 193 | 5925 | | |
| 195 | 194 | 5739 | | |
| 196 | 195 | 6650 | | |
| 197 | 196 | 6304 | | |
| 198 | 197 | 6227 | | |
| 199 | 198 | 7211 | | |
| 200 | 199 | 6665 | | |
| 201 | 200 | 6031 | | |

Figure 1.12: Coloriage des déciles inférieur et supérieur d'une série de 200 valeurs de productions laitières annuelles.

La production laitière de 200 vaches a été mesurée et mémorisée dans une feuille Excel (les numéros de vache sont dans la première colonne et les productions laitières en kilos de lait dans la seconde). On demande de marquer en vert le décile supérieur (vaches à utiliser pour la reproduction) et en rouge le décile inférieur (vaches à réformer).

▶

Les données pour ce type de problèmes proviennent généralement de données mesurées sur le terrain par l'expérimentateur. Comme nous ne disposons pas ici de ce type de données, nous allons simuler cette expérience en générant les données de manière aléatoire. On commence par mettre les en-têtes de colonnes "Vache" et "Kg Lait" respectivement en A1 et B1. Les numéros des vaches de 1 à 200 seront ensuite introduits dans la plage A2:A201, come expliqué plus haut. En ce qui concerne les productions, nous allons employer une procédure dont le fonctionnement sera expliqué plus tard (voir le chapitre sur l'échantillonnage). On procède comme suit:

- dans la cellule B2, on tape:

  $$= ARRONDI(LOI.NORMALE.INVERSE(ALEA(); 6000; 1000); 0)$$

  Cette formule tire une valeur au hasard dans une distribution de moyenne

6000 et de déviation standard 1000 (censée simuler la distribution de la production laitière) et arrondit le résultat à 0 décimale.

- On copie cette formule dans les cellules B3:B201 pour obtenir 200 échantillons aléatoires.
- Comme les productions (cellules B2:B201) varient à chaque action sur la feuille, on les "fige" de la manière suivante: on sélectionne la plage B2:B201, on "copie" cette zone, puis on fait un "collage spécial" sur la même zone. pour cela, on clique sur le bouton de droite après avoir "copié", on choisit l'option "Collage spécial" dans le menu qui apparait, et on sélectionne l'option "Valeurs" dans le menu qui apparait. On valide ensuite ce choix en appuyant sur "Ok". On peut vérifier que les formules de la plage B2:B201 ont disparu, et que seules les valeurs ont été conservées.

La figure 1.12 montre le début et la fin du fichier utilisé. On voit que les productions à utiliser sont dans la plage B2:B201. On peut utiliser la fonction CENTILE vue plus haut pour obtenir les déciles inférieur (cellule D4) et supérieur (cellule D6). Le "coloriage" des cellules peut ensuite se faire facilement avec une macro:

```
Sub Deciles()
 For l = 2 To 201
  If (Cells(l, 2).Value < Range("D4").Value) Then
   Cells(l, 2).Interior.ColorIndex = 3
  End If
  If (Cells(l, 2).Value > Range("D6").Value) Then
   Cells(l, 2).Interior.ColorIndex = 4
  End If
 Next l
End Sub
```

◀

## 1.2 Introduction R

### 1.2.1 Prérequis

Les explications sur le logiciel R figurant dans le cours "l'ABC d'R" disponible sur le site du cours de biostatistique, nous nous limiterons ici à quelques rappels. Par conséquent, le parcours du site est conseillé et sera considéré comme un prérequis pour effectuer les exercices proposés dans cette partie.

### 1.2.2 Données (scalaires et vecteurs - séquences)

En R, on représente les données par des variables. Ces variables ont un nom, un type et une valeur (et éventuellement d'autres propriétés).
Le nom, à quelques contraintes près (comme ne pas mettre certains caractère spéciaux, ou ne pas employer de mots-clés du langage), sont quelconques et peuvent être choisis par l'utilisateur. Il faut signaler que la casse utilisée est importante: ainsi, par exemple, ab, AB, Ab et aB pourraient désigner 4 variables

distinctes.
Le type de variables désigne le type d'information que la variable est susceptible de contenir. Une variable de type scalaire permet de mémoriser une seule valeur (numérique ou alphabétique). Une variable de type vectoriel permet de mémoriser plusieurs valeurs. Une variable de type matriciel permet de mémoriser un tableau de valeurs (dans ces notes, le plus souvent, le tableau aura 2 dimensions, mais ce n'est pas une obligation). Une variable de type dataframe ressemble à une matrice, à la différence près que les colonnes du tableau peuvent contenir des informations de natures différentes, alors que toutes les colonnes d'un tableau contiennent des information de même nature (par exemple, des nombres entiers).
Dans ce qui suit, nous allons illustrer comment il est possible de mémoriser des nombres dans des variables, nombres que nous pouvons ensuite utiliser pour des calculs ultérieurs. Nous commencerons par le cas de variables scalaires

```
> # Donner une valeur aux variables scalaires a et b
> a<-10
> b<-2.2
> # Afficher le contenu de a
> a
[1] 10
> # Utiliser le contenu des variables
> x<-3
> y<-a+b*x
> y
[1] 16.6
```

Passons à présent au cas des variables vectorielles:

```
> # Creer explicitement un vecteur de 5 valeurs numeriques
> # Une valeur par defaut (0, ici) est memorisee pour chaque element
> v<-vector(mode="numeric",length=5)
> # Creer implicitement un vecteur de 3 valeurs numeriques
> # Les valeurs sont memorisees des la creation
> w<-c(10,20,30)
> # Afficher le contenu des vecteurs
> v
[1] 0 0 0 0 0
> w
[1] 10 20 30
> # Acceder aux elements du vecteur
> w[1]
[1] 10
> v[1]<-5
> v[2]<-10
> v
[1] 5 10 0 0 0
> v[c(1,2)]
[1] 5 10
> # Utiliser le contenu des vecteurs
> droite<-c(10,2.2)
> x<-c(1,2,3,4,5)
> y<-droite[1]+droite[2]*x
> y
[1] 12.2 14.4 16.6 18.8 21.0
```

Lorsque des vecteurs correspondent à des séquences de nombres, il y a des possibilités pour faciliter la création de ces vecteurs:

```
> # Creer un vecteur des 100 premiers entiers
> z100<-1:100
> # Afficher les 10 premiers elements du vecteur
> z100[1:10]
[1] 1 2 3 4 5 6 7 8 9 10
> # Creer un vecteur des multiples de 3 inferieurs a 20
> m3<-seq(0,20,3)
> m3
[1] 0 3 6 9 12 15 18
> z100[m3]
[1] 3 6 9 12 15 18
# Remarquez qu'il n'y a pas d'element d'indice 0
# Creer un vecteur de valeurs identiques (10, ici)
> vec<-rep(10,5)
> vec
[1] 10 10 10 10 10
```

Une matrice est un tableau de nombres de n lignes et m colonnes, où n et m doivent être précisés. Un exemple de création de matrices est donné ci-dessous:

```
> # Creer une matrice de 10 lignes et 5 colonnes
> # contenant les valeurs de 1 a 50.
> # L'option nr=10 precise qu'il y aura 10 rangees (et donc 5 colonnes↩
    )
> # L'option byrow=TRUE precise que les 50 valeurs seront placees
> # dans la matrice rangee par rangee
> XX<-matrix(1:50,nr=10,byrow=TRUE)
> # Afficher la matrice
> XX
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1    2    3    4    5
 [2,]    6    7    8    9   10
 [3,]   11   12   13   14   15
 [4,]   16   17   18   19   20
 [5,]   21   22   23   24   25
 [6,]   26   27   28   29   30
 [7,]   31   32   33   34   35
 [8,]   36   37   38   39   40
 [9,]   41   42   43   44   45
[10,]   46   47   48   49   50
> # Afficher des elements de la matrice:
> # l'element sur la 2eme ligne et la 3eme colonne
> XX[2,3]
> # la 4eme colonne
> XX[,4]
[1]  4  9 14 19 24 29 34 39 44 49
> # les 2 premiers elements de la 4eme colonne
> XX[1:2,4]
[1] 4 9
> les premiers et derniers elements des 2 dernieres colonnes
> XX[c(1,10),4:5]
     [,1] [,2]
[1,]    4    5
[2,]   49   50
```

Le type dataframe, évoqué plus haut, sera montré après que la lecture de données ait été introduite.

### 1.2.3 Fonctions

Comme Excel, R est doté d'une riche panoplie de fonctions. Une fonction de R est, comme en Excel, caractérisée par un nom suivi d'une liste d'arguments

séparés par une virgule (et non un point-virgule comme dans Excel) placée entre parenthèses. Dans le listing qui suit, on va illustrer l'utilisation de quelques fonctions utiles, mais ce n'est qu'un aperçu très sommaire des possibilités offertes.

```
> # Creation d'un vecteur des 100 premiers nombres
> v<-1:100
> m<-mean(v)
> s<-sum(v)
> m-s/100
[1] 0
> v<-var(v)
> # La deviation standard est la racine carree de la variance
> # La racine carree se dit "square root" en anglais
> ds<-sqrt(v)
> etendue<-max(v)-min(v)
> etendue
[1] 99
> range(v)
[1] 1 100
> median(v)
[1] 50.5
> quartiles<-quantile(v,c(.25,.50,.75))
> quartiles
  25%   50%   75%
25.75 50.5 75.25
> deciles<-quantile(v,seq(0.1,0.9,0.1))
 10%  20%  30%  40%  50%  60%  70%  80%  90%
10.9 20.8 30.7 40.6 50.5 60.4 70.3 80.2 90.1
```

### 1.2.4 Lecture de données

Dans certains cas, les données utilisées dans R auront été générées dans le logiciel, comme sur les petits exemples qui précèdent. Dans d'autres, les données viendront de sources externes. Ce sera par exemple le cas lorsqu'on souhaite analyser les données récoltées dans une expérience et que ces données ont été rassemblées dans un fichier texte ou un fichier Excel. Dans ce cas, pour effectuer les manipulations dans R, il y aura une étape préalable d'importation des données dans le logiciel. R dispose de nombreuses possibilités permettant ces importations, nous nous limiterons ici au cas de données en mode texte (pour importer des données depuis une feuille Excel, il est possible de tout d'abord sauvegarder le document Excel en texte délimité par des tabulations ou un autre caractère, puis d'importer le fichier texte dans R). Le principe est d'employer une fonction de lecture, qui prendra comme arguments le fichier dans lequel les données se trouvent et d'autres paramètres permettant de préciser certaines options de lecture, et qui retournera un objet qu'on pourra mémoriser, puis manipuler dans l'environnement de R. Le type de l'objet qui est retourné est "dataframe". Si les données peuvent être vues comme un ensemble de m mesures (comme l'âge, le sexe, la taille, le poids, etc...) prises sur n individus, le dataframe peut être vu comme un tableau de n lignes (correspondant aux n observations) et m colonnes (correspondant aux m attributs mesurés pour chaque observation). Contrairement aux matrices, les données de cellules différentes peuvent donc être de types différents. Illustrons maintenant la manière dont cette lecture de fichiers fonctionne. Pour l'exemple, nous supposerons qu'un fichier (dont un extrait est reproduit ci-dessous) a été déposé dans le répertoire de travail de R (un répertoire que l'on peut choisir avec le menu "Fichier" du logiciel):

| Chien | Sexe | Taille | Poids | Robe |
|---|---|---|---|---|
| Leila | F | 40 | 21.2 | Dorée |
| Bill | M | 38 | 22.0 | Dorée |
| Lolita | F | 34 | 16.6 | Brune |
| Max | M | 42 | 23.1 | Brune |
| ... | ... | ... | ... | ... |
| Bobby | M | 38 | 24.5 | Dorée |

On supposera que ce fichier est appelé 'chiens.txt', et que les différents champs de chaque ligne sont séparés dans le fichier par des tabulations. La lecture du fichier par R se fera de la manière suivante:

```
> f<-read.table(file="chiens.txt",sep="\t",head=TRUE,dec=".")
> f
   Chien  Sexe  Taille  Poids   Robe
1   Leila     F      40   21.2  Doree
2    Bill     M      38   22.0  Doree
3  Lolita     F      34   16.6  Brune
4     Max     M      42   23.1  Brune
5   Laika     F      35   18.8  Doree
6   Simba     M      46   25.0  Doree
7   Vodka     M      39   20.4  Brune
8   Pixel     M      34   18.1  Brune
9   Bobby     M      38   24.5  Doree
```

La première instruction est l'instruction de lecture du fichier. Comme cette fonction comporte plusieurs arguments, et qu'il n'est pas facile de se souvenir de l'ordre dans lequel il faut les lui soumettre, R permet d'utiliser des arguments nommés, prenant la forme:

$$<\text{argument}> = <\text{valeur}>$$

L'ordre des arguments est alors quelconque. Détaillons à présent les 4 arguments qui ont été fournis:

- file="chiens.txt" permet de spécifier le nom du fichier (texte) à utiliser,
- $sep =$ "$\backslash t$" sert à préciser que les champs dans le fichier sont séparés par des tabulations (le symbole "$\backslash t$" représente une tabulation). Ce paramètre sert donc à préciser le caractère de séparation des champs (dans certains formats, ce caractère pourrait être "," ou ";"),
- $head = TRUE$ est utilisé pour signaler qu'il y a une ligne en tête de fichier qui n'est pas une ligne de données, mais bien une ligne de titre. Si une telle ligne n'était pas présente, on spécifierait $head = FALSE$,
- $dec =$ "." précise que le séparateur décimal est un "." (et non pas une ",").

Les données sont donc lues dans une variable, appelée f dans l'exemple, qui est de type "dataframe". La deuxième instruction permet d'afficher le fichier. D'autres instructions sont disponibles pour obtenir le nom des variables constituant le dataframe, et accéder aux valeurs. Quelques exemples sont donnés dans le listing qui suit:

```
> # Creation du dataframe
> f<-read.table(file="chiens.txt",sep="\t",head=TRUE,dec=".")
> # Quels sont les noms des variables du dataframe ?
> names(f)
[1] "Chien"  "Sexe"   "Taille" "Poids"  "Robe"
> # Afficher les 3 premieres lignes du dataframe
> head(f,n=3)
   Chien Sexe Taille Poids  Robe
1  Leila    F     40  21.2 Doree
2   Bill    M     38  22.0 Doree
3 Lolita    F     34  16.6 Brune
> # Afficher les 2 dernieres lignes du dataframe
> tail(f,n=2)
8  Pixel    M     34  18.1 Brune
9  Bobby    M     38  24.5 Doree
> # Afficher le nombre de lignes du dataframe
> nrow(f)
[1] 9
> # Afficher les tailles
> f$Taille
[1] 40 38 34 42 35 46 39 34 38
> # Afficher les tailles des 3 premiers chiens
> f$Taille[1:3]
[1] 40 38 34
> # Afficher les tailles des femelles
> f$Taille[f$Sexe=="F"]
[1] 40 34 35
```

La dernière instruction nécessite sans doute quelques explications. L'expression $f\$Sexe == "F"$ est un exemple d'expression logique, c'est-à-dire d'expression dont le résultat est "vrai" ou "faux". Le symbole " == " est le symbole d'égalité dans les expressions logiques: l'expression est "vraie" si ce qui figure à gauche du symbole est égal à ce qui figure à droite du symbole. Dans l'exemple, ce qui figure à gauche est un vecteur (contenant les valeurs représentant le sexe des différents individus de l'expérience) et ce qui figure à droite est un scalaire (la lettre "F"). Quand on demande à R de comparer un vecteur à un scalaire, R compare chaque élément du vecteur au scalaire, et pour chaque comparaison, retourne TRUE si l'expression est vraie et FALSE dans le cas contraire. Le résultat est donc un vecteur de valeurs logiques. On peut voir ce vecteur:

```
> f$Sexe=="F"
[1]  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
```

Enfin, si on prend ce vecteur logique pour indexer un vecteur, seules les coordonnées pour lesquelles la valeur logique est TRUE est retournée (dans notre exemple, la première, la troisième et la cinquième). Cette possibilité facilite évidemment les calculs quand on souhaite, par exemple, calculer le poids moyens des chiens dorés ou la taille des femelles.

### 1.2.5 Instructions for et if

Encore plus qu'Excel, R est un véritable langage informatique, permettant de faire des manipulations simples ou complexes sur les jeux de données. A ce titre, il est doté notamment d'instructions permettant de faire des boucles ou de faire des choix conditionnels, fonctionnalités présentes également dans les macros d'Excel. Commençons par les boucles. Plusieurs instructions existent (for, repeat, while, etc...), nous verrons essentiellement la boucle for, les autres

ayant un comportement similaire. La syntaxe est la suivante:

$$for(< variable > in < vector >)\{< instructions >\{$$

$< variable >$ désigne le nom d'une variable, qui prendra successivement chaque valeur du vecteur $< vecteur >$. Pour chaque valeur de $< variable >$, les instructions $< instructions >$ seront exécutées. L'exemple qui suit illustre l'utilisation de cette structure:

```
> # Calcul de la somme des 100 premiers nombres et de leurs carres
> s<-0
> s2<-0
> for (i in 1:100) { s<-s+i; s2<-s2+i*i }
> s
[1] 5050
> s2
[1] 338350
> # Autre calcul de ces sommes
> l<-1:100
> sum(l)
[1] 5050
> sum(l*l)
[1] 338350
> # Calcul utilisant les formules
> # s<-n*(n+1)/2
> # s2<-n*(n+1)*(2*n+1)/6
> 100*101/2
[1] 5050
> 100*101*201/6
[1] 338350
```

L'exécution conditionnelle d'instructions se fait avec l'instruction:

$$if < expression > then\{< instr_A >\} else\{< instr_B >\}$$

où $< expression >$ est une expression logique, c'est-à-dire dont l'évaluation retourne TRUE si l'expression est vraie, et FALSE dans le cas contraire (voir plus haut pour un exemple). Si l'évaluation retourne TRUE, le(s) instruction(s) $< inst_A >$ sont effectuées. Dans le cas contraire, le(s) instruction(s) $< instr_B >$ sont effectuées. Dans le cas où il n'y a aucune instruction à effectuer quand l'évaluation de l'expression logique retourne FALSE, la partie $else < instr_B >$ peut être omise. L'exemple suivant illustre l'utilisation de cette instruction:

```
> # Calcul des poids moyens de males et des femelles
> pm<-0; pf<-0; nm<-0; nf<-0
> # Boucle sur les rangees du dataframe
> for (i in 1:nrow(f)) {
+  # Test sur le sexe
+  if (f$Sexe[i]=="M") { pm<-pm+f$Poids[i]; nm<-nm+1 }
+                 else { pf<-pf+f$Poids[i]; nf<-nf+1 }
+ }
> pm/nm
[1] 22.18333
> pf/nf
[1] 18.86667
> # Alternativement:
> mean(f$Poids[f$Sexe=="M"])
[1] 22.18333
> mean(f$Poids[f$Sexe=="M"])
[1] 18.86667
```

Quelques remarques sur ce code s'imposent:

- si on souhaite écrire plusieurs instructions sur la même ligne, comme pour la seconde ligne du code, il faut séparer les instructions successives par un ";"
- on peut intercaler des espaces pour la lisibilité, comme par exemple pour la ligne où figure le "else",
- il est parfois utile, toujours pour des questions de lisibilité, d'écrire certaines instructions (comme les instructions "for" et "if" dans le code) sur plusieurs lignes. Lors du passage à la ligne, si l'instruction n'est pas terminée, R remplace son caractère de sollicitation (le ">" en début de ligne) par un autre caractère (le "+" qui apparait sur les lignes concernées). Une fois l'instruction terminée, elle est exécutée et le ">" réapparait.

### 1.2.6 Utilisation de la fenêtre de script

On voit, dans les exemples précédents, qu'on peut être amené à construire de petits programmes (appelés 'scripts' dans R) pour obtenir les résultats recherchés. Plutôt que de taper toutes les commandes dans la console de R, au risque de devoir les retaper à plusieurs reprise pour corriger les erreurs, il est parfois plus simple d'introduire les commandes (c'est-à-dire le 'script') dans une fenêtre qui ressemble à un éditeur de texte, puis de copier-coller ces commandes dans la console pour en vérifier le bon déroulement. En cas d'erreur, il suffit de modifier la commande fautive dans la fenêtre de l'éditeur contenant le script, et de recommencer le copier-coller. De plus, les scripts ainsi créés peuvent être sauvegardés pour être récupérés lors d'une session ultérieure. La procédure est la suivante: on commence par créer un nouveau script (voir la figure 1.13). Une fois la commande validée, une fenêtre s'ouvre, dans laquelle on peut taper le code (voir la figure 1.14) et le copier-coller dans la console ou, alternativement, sauvegarder le script dans le répertoire de travail de R via le menu File. Vous pourrez alors exécuter les lignes de commandes du script depuis la console R grâce à l'instruction source(« nom de votre script avec son extension »).

**Exercice 1.2.1**

Sachant que la fonction rnorm(25,mean=100,sd=10) permet de générer 25 valeurs aléatoirement dans une distribution normale de moyenne $\mu = 100$ et de déviation standard $\sigma = 10$, calculez les moyennes pour 1000 échantillons de ce type. Calculez ensuite la moyenne et la déviation standard de ces 1000 moyennes et commentez votre résultat.

►

**Solution**: les résultats obtenus devraient permettre d'illustrer le théorème de la limite centrale: la moyenne des moyennes devrait être proche de la moyenne $\mu$ des observations, et la déviation standard devrait être proche de sa valeur théorique $\sigma/\sqrt{n}$. De plus, la distribution des moyennes devrait être proche d'une distribution normale.

◄

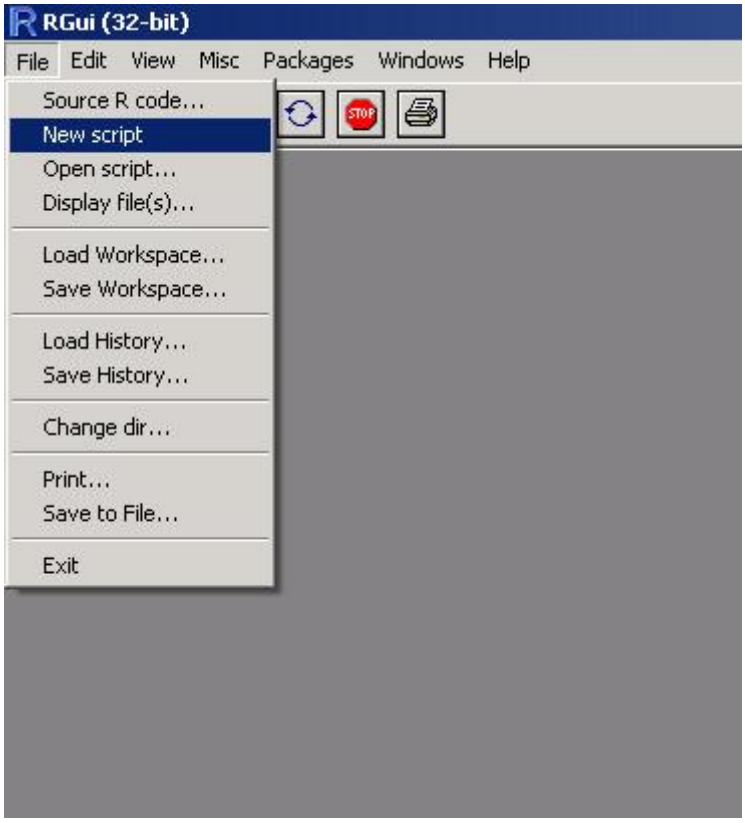



Figure 1.13: Ouverture d'une nouvelle fenêtre de script.

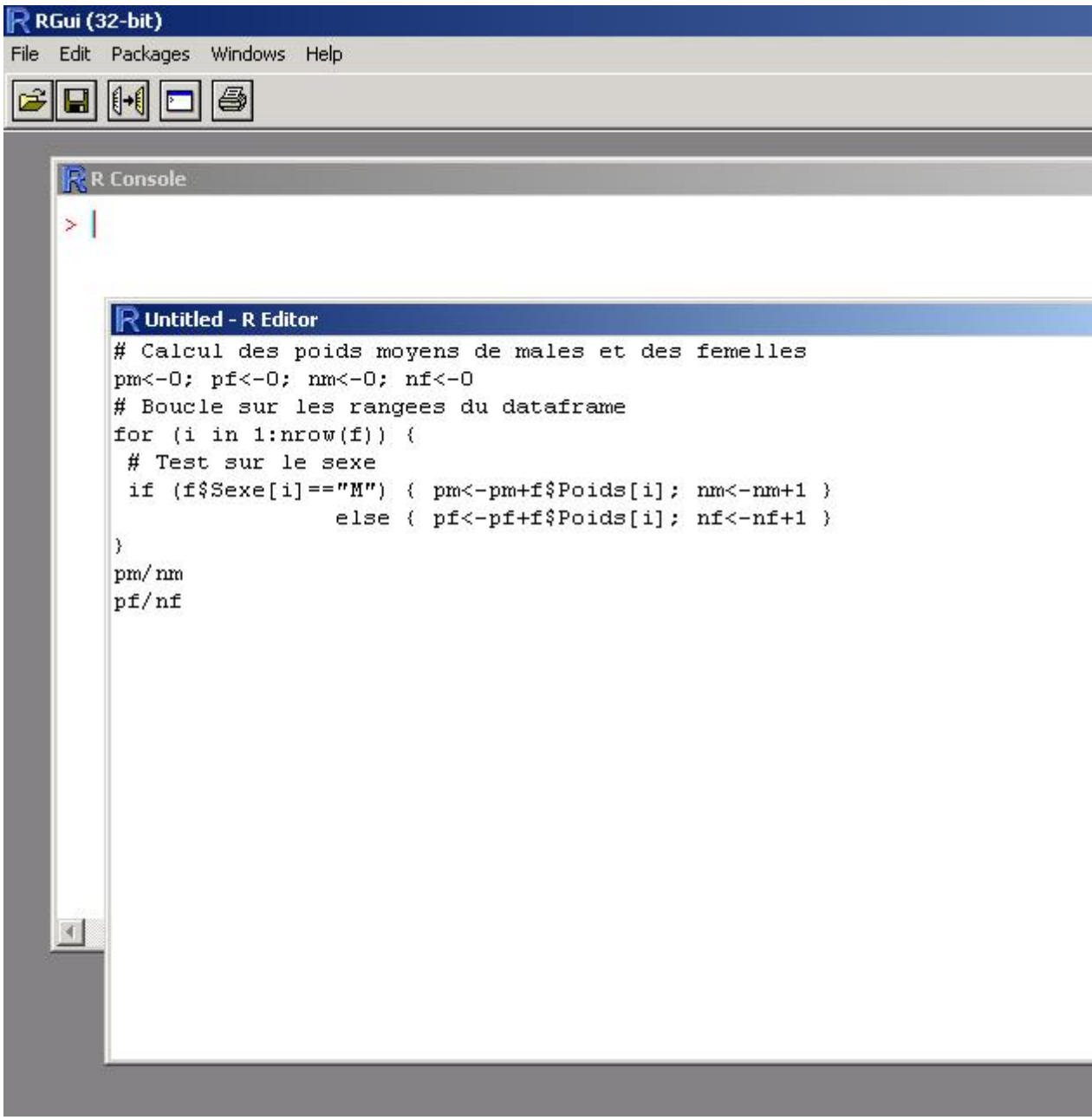



Figure 1.14: Utilisation de la fenêtre de script.

## 1.3 Distributions et calcul de probabilités(Séance TP 2 sauf le paragraphe 1.3.3)

Excel et R sont particulièrement adaptés pour les calculs de probabilités faisant appel aux distributions (celles vues au cours ainsi que d'autres). Dans ce qui suit, nous donnerons quelques exemples de calculs résolus "manuellement", puis nous utiliserons les deux logiciels afin de démontrer la relative facilité de l'exercice.

### 1.3.1 Loi binomiale

**Prérequis**

- conditions d'utilisation de la distribution binomiale
- formule de la distribution binomiale

**Exercice résolu 1.3.1**

Un généticien souhaite tester 10 vaches pour voir si elles sont homozygotes XX pour une mutation qui affecte la production de lait. Il a été estimé que la proportion de vaches de ce type dans la population visée est de l'ordre de 30 %. Quelle est la probabilité d'avoir 3 vaches qui présente ce génotype dans l'échantillon ? Et quelle est la probabilité d'en avoir plus de trois ?

▶

Il s'agit d'un calcul qui respecte les conditions d'utilisation de la loi binomiale (le résultat du tirage est binaire (l'individu est de génotype XX ou pas), le nombre de tirages est fixé ($n = 10$), la probabilité d'obtenir un génotype XX à un tirage particulier est fixe (tirage avec remise, pour lequel $p = 0.30$) et les tirages sont indépendants (on tire au hasard dans une population de vaches, et les individus échantillonnés successivement sont a priori non apparentés)). Par conséquent, on peut utiliser la formule de la distribution binomiale pour calculer la probabilité associée au fait d'obtenir $r$ fois le génotype XX lors de $n$ tirages (avec, évidemment, $0 \leq r \leq n$)

- Solution "manuelle"
  Pour rappel, la formule de la distribution binomiale est:

$$P(r|n,p) = \frac{n!}{r! * (n-r)!} * p^r * (1-p)^{(n-r)}$$

  Dans le problème qui nous intéresse, on peut donc calculer facilement la probabilité d'obtenir exactement 3 génotypes XX si on teste 10 vaches:

$$\begin{aligned} P(3|n=10,p=0.3) &= P(3|10,0.3) && (1.1)\\ &= \frac{10!}{3! * 7!} * 0.3^3 * 0.7^7 && (1.2)\\ &= 120 * 0.027 * 0.0823543 && (1.3)\\ &= 0.26683 && (1.4) \end{aligned}$$

La probabilité P d'avoir plus de 3 homozygotes GG se calcule en additionnant les situations correspondantes, soit:

$$\begin{aligned} P &= P(4|10,0.3) + P(5|10,0.3) + \cdots + P(10|10,0.3) && (1.5)\\ &= 1 - P(0|10,0.3) - P(1|10,0.3) - P(2|10,0.3) - && (1.6)\\ &\quad 1 - 0.02825 - 0.12106 - 0.23347 - 0.26683 && (1.7)\\ &= 0.35039 && (1.8) \end{aligned}$$

- Solution "Excel" On utilise pour ce type de calculs, la fonction LOI.BINOMIALE, qui a 4 arguments:
  1. le nombre de fois où l'événement d'intérêt (ici, obtenir un génotype XX) se produit (r),
  2. le nombre de tirages (n),
  3. la probabilité associée à l'événement d'intérêt à chaque tirage (p),
  4. une valeur booléenne (VRAI ou FAUX). Si on indique VRAI, la probabilité calculée est la somme des probabilités associées à $r, r-1, \cdots, 1, 0$ (probabilité cumulée). Si on indique FAUX, la probabilité calculée est celle associée à r seul (probabilité simple).

  La réponse à la première question s'obtient donc via la formule:

$$= LOI.BINOMIALE(3; 10; 0.3; FAUX)$$

  qui fournit la valeur 0.26683. La réponse à la seconde question se calcule via:

$$= 1 - LOI.BINOMIALE(3; 10; 0.3; VRAI)$$

  qui fournit la valeur 0.35039.

- Solution "R" Les fonctions de R qui permettent de faire des calculs avec la distribution binomiale se terminent toutes par *binom*. Celle qui permet de calculer des probabilités simples est:

$$dbinom(x = r, size = n, prob = p)$$

  (voir les remarques sur les arguments des fonctions R). Pour le problème qui nous intéresse ici, on peut obtenir le premier résultat en tapant:

```
> dbinom(3,10,0.3)
[1] 0.2668279
```

  La fonction qui permet le calcul de probabilités cumulées est:

$$pbinom(x = r, size = n, prob = p, lower.tail = TRUE)$$

  Le dernier argument permet de dire si on veut calculer $P(x \leq r)$ (en indiquant $lower.tail = TRUE$) ou $P(x > r)$ (en indiquant $lower.tail = FALSE$). Par défaut (c'est-à-dire si on indique rien), l'option $lower.tail = TRUE$ est utilisée. On peut donc répondre à notre question en employant les commandes suivantes de R:

```
> 1-pbinom(3,10,0.3)
[1] 0.3503893
> pbinom(3,10,0.3,lower.tail=FALSE)
[1] 0.3503893
```

◀

**Exercice 1.3.2**

Pour effectuer des mesures immunologiques sur un échantillon suffisamment grand (pour des raisons statistiques) mais pas trop grand (pour des raisons de coûts), un vétérinaire souhaiterait obtenir entre 20 et 50 animaux ayant un statut immunitaire particulier. Dans la population visée, 20 % des individus ont ce statut. Il décide, pour constituer son échantillon de travail, de partir d'un échantillon aléatoire de 100 animaux, espérant qu'il obtiendra dans cet échantillon entre 20 et 50 individus avec le statut recherché. Quelle est la probabilité qu'il obtienne effectivement ce résultat ? Faites le calcul avec Excel et avec R, et vérifiez que les deux approches fournissent le même résultat.

▶

**Solution**: la probabilité vaut *0.5398386*.

◀

**Exercice 1.3.3**

Dans le même problème, combien d'animaux doit-il échantillonner au minimum pour avoir une probabilité supérieure à 80% d'atteindre son objectif ?

▶

**Solution**: l'effectif minimal est *116*. Avec cet effectif, la probabilité d'avoir entre 20 et 50 individus de statut recherché est 0.80313.

◀

### 1.3.2 Loi de Poisson

**Prérequis**

- conditions d'utilisation de la distribution de Poisson
- formule de la distribution de Poisson
- conditions d'équivalence entre une distribution de Poisson et une autre distribution

**Exercice résolu 1.3.4**

La peste porcine africaine (PPA) frappe les exploitations d'une région infectée, à raison en moyenne de 3 nouvelles exploitations par semaine. En cette période épidémique, quelle est la probabilité que 3 exploitations soient frappées la même semaine ? Et qu'il y en ait plus de 5 ?

▶

Dans ce problème, les conditions d'utilisation d'une loi de Poisson sont rencontrées: événement binaire (contamination/pas de contamination) dont on comptabilise le nombre d'occurrences sur des exploitations indépendantes, dont le nombre n'est à priori pas limité (on considère le nombre d'exploitations comme très grand).

- Solution "manuelle"
  La formule de la distribution de Poisson permettant de calculer la probabilité d'occurrence de r événements (ici, un "événement" est la découverte d'une nouvelle exploitation contaminée par la PPA) est:

$$P(r|\mu) = \frac{e^{-\mu} * \mu^r}{r!}$$

  où $\mu$ est le nombre moyen d'occurrences de l'événement sur la période d'étude. Dans notre problème, il y a en moyenne 3 occurrences par semaine, et donc la valeur de $\mu$ est 3. Le premier calcul est alors une simple application de la formule:

$$\begin{aligned} P(r=3|\mu=3) &= \frac{e^{-3} * 3^3}{3!} && (1.9)\\ &= \frac{4.5}{e^3} && (1.10)\\ &= 0.224 && (1.11) \end{aligned}$$

  Pour le calcul de $P = P(r > 5)$, il faudrait calculer:

$$P = P(r=6|\mu=3) + P(r=7|\mu=3) + \cdots$$

  La difficulté est évidemment qu'il y a une infinité de termes dans ce calcul. On peut remplacer ce calcul en notant que:

$$\begin{aligned} &P(r=0|\mu=3) + P(r=1|\mu=3) + \cdots + P(r=5|\mu=3) + P(r=6|\mu=3) + \\ &P(r=7|\mu=3) + \cdots = 1 \end{aligned}$$

  et donc que:

$$\begin{aligned} &P(r=6|\mu=3) + P(r=7|\mu=3) + \cdots = \\ &1 - P(r=0|\mu=3) - P(r=1|\mu=3) - \cdots - P(r=5|\mu=3) = P \end{aligned}$$

  Par conséquent, le calcul à effectuer est:

$$P = 1 - e^{-3} * (\frac{3^0}{0!} - \frac{3^1}{1!} - \frac{3^2}{2!} - \frac{3^3}{3!} - \frac{3^4}{4!} - \frac{3^5}{5!})$$

  ce qui fournit:

$$P = 1 - \frac{1+3+4.5+4.5+3.375+0.025}{e^3} = 1 - 0.9160821 = 0.083918$$

- Solution "Excel"
  La fonction LOI.POISSON est utilisée pour les calculs sur cette distribution. Cette fonction prend 3 paramètres:

  1. le nombre de fois où l'événement d'intérêt (ici, où une exploitation est trouvée positive) se produit (r),
  2. le nombre moyen d'événements par unité de temps ($\mu$, qui est ici le nombre moyen d'exploitations trouvées positives par semaine, soit $\mu = 3$),

3. une valeur booléenne (VRAI ou FAUX). Si on indique VRAI, la probabilité calculée est la somme des probabilités associées à $r, r-1, \cdots, 1, 0$ (probabilité cumulée). Si on indique FAUX, la probabilité calculée est celle associée à r seul (probabilité simple).

Pour répondre à la première question, il suffit donc de taper la formule:

$$= LOI.POISSON(3; 3; FAUX)$$

qui fournit la valeur 0.22404181. La réponse à la seconde question se calcule via:

$$= 1 - LOI.POISSON(5; 3; VRAI)$$

qui fournit la valeur 0.083918.

- Solution "R"
  Les fonctions de R qui permettent de faire des calculs avec la distribution de Poisson se terminent toutes par *pois*. Celle qui permet de calculer des probabilités simples est:

$$dpois(x = r, lambda = \mu)$$

(voir les remarques sur les arguments des fonctions R). Pour le problème qui nous intéresse ici, on peut obtenir le premier résultat en tapant:

```
> dpois(3,3)
[1] 0.2240418
```

Pour le calcul de probabilités cumulées, on emploie:

$$ppois(x = r, lambda = \mu, lower.tail = TRUE)$$

Le dernier argument permet de dire si on veut calculer $P(x \leq r)$ (en indiquant $lower.tail = TRUE$) ou $P(x > r)$ (en indiquant $lower.tail = FALSE$). Par défaut (c'est-à-dire si on indique rien), l'option $lower.tail = TRUE$ est utilisée. On peut donc répondre à notre question en employant indifféremment les commandes suivantes de R:

```
> 1-ppois(5,3)
[1] 0.08391794
> ppois(5,3,lower.tail=FALSE)
[1] 0.08391794
```

◀

**Exercice 1.3.5**

Si le nombre de saisies par jour dans un abattoir est 5 en moyenne, quelle est la probabilité d'avoir 7 saisies en 2 jours ? Résolvez cet exercice soit en raisonnant sur une distribution du nombre de cas par 2 jours, soit en combinant les résultats obtenus chacun des deux jours (avec la distribution du nombre de cas par jour) et en combinant les résultats, et montrez que les deux approches conduisent au

même résultat.

▶

**Solution**: la probabilité vaut 0.09007

◀

**Exercice 1.3.6**

Lors d'un dénombrement bactérien, après avoir dilué 5 fois la solution initiale (à chaque dilution, la concentration devient le dixième de ce qu'elle était à l'étape précédente), le bactériologiste constate que les sur les 30 boites de Pétri ensemencées avec 1 ml de la solution contenant les bactéries, 10 sont ne présentent aucune plage de lyse. Quelle est la concentration moyenne en bactérie du liquide initial ?

▶

**Solution**: la concentration est de $1.0986 * 10^8$ bactéries/litre.

◀

### 1.3.3 Loi de $\chi^2$ (Séance TP 5)

**Prérequis**

- principes des tests d'hypothèses
- tests d'association entre variables discrètes
- conditions d'utilisation de la distribution de $\chi^2$
- formule du $\chi^2$

**Exercice résolu 1.3.7**

Il existe 3 robes majeures chez le chien labrador: dorée (60% de la population), noire (30% de la population) et chocolat (10% de la population). Un vétérinaire orthopédiste s'intéresse au lien éventuel entre la robe et les problèmes de dysplasie de la hanche, problème fréquent dans cette race. Après consultation des fiches cliniques dans plusieurs cliniques de la région où il travaille, il a fait le relevé suivant des cas de dysplasie:

| Couleur | Chocolat | Dorée | Noire | Total |
|---|---|---|---|---|
| Nombre de cas | 31 | 80 | 48 | 159 |

Sur base de ces données, peut il conclure à une association entre la robe et les problèmes de dysplasie ?

▶

Ce type de problèmes est clairement un test d'hypothèse: le vétérinaire souhaite tester son hypothèse selon laquelle l'apparition de problèmes de dysplasie est liée à la robe. Plus précisément, il pense que certaines robes prédisposent plus à l'apparition de ces problèmes de dysplasie. L'approche pour tester cette hypothèse est l'approche classique utilisée en statistiques: on se pose la question de savoir s'il est probable d'obtenir de tels résultats s'il n'y avait en réalité pas

d'effet de la robe sur l'apparition de dysplasies. Plus formellement, on pose l'hypothèse nulle selon laquelle, la prévalence est la même pour chaque robe:

$$H_0 : p(d|C) = p(d|D) = p(d|N) = p(d)$$

où les lettres C, D et N sont les initiales des robes correspondantes et la lettre d est l'initiale de dysplasie. Si l'hypothèse nulle est vraie:

$$p(D|d) = \frac{p(D,d)}{p(d)} = \frac{p(D) * p(d|D}{p(d)} = \frac{p(D) * p(d)}{p(d)} = p(D)$$

De manière similaire, on déduirait que $p(C|d) = p(C)$ et $p(N|d) = p(N)$. Par conséquent, on peut donner les effectifs auxquels il faut s'attendre dans notre échantillon d'animaux dysplasiques:

| Couleur | Chocolat | Dorée | Noire |
|---|---|---|---|
| Nombre de cas observés | 31 | 80 | 48 |
| Nombre de cas attendus | 0.1*159 = 15.9 | 0.6*159 = 95.4 | 0.3*159 = 47.7 |

La comparaison des observés et des attendus se fait classiquement en calculant la valeur de $\chi^2$ avec la formule:

$$\chi^2 = \sum_i \frac{(O_i - A_i)^2}{A_i}$$

où $O_i$ et $A_i$ désignent les observés et les attendus de la catégorie i, respectivement. Pour l'exemple qui nous occupe:

$$\chi^2 = \frac{(31 - 15.9)^2}{15.9} + \frac{(80 - 95.4)^2}{95.4} + \frac{(48 - 47.7)^2}{47.7} = 16.828$$

Comme notre problème comporte 3 catégories, le nombre de degrés de liberté à considérer est $(3 - 1) = 2$. La probabilité de s'écarter aussi fort, voire plus fort, de ce qui était prévu peut donc s'obtenir en calculant la probabilité qu'une valeur de $\chi^2$ prise dans une distribution avec 2 degrés de liberté soit supérieure ou égale à 16.828. Ce calcul nécessite de calculer l'intégrale correspondante, ce qui n'est pas simple... Une alternative est de consulter une table de valeurs seuil des distributions de $\chi^2$ (par exemple dans le syllabus). On constate alors qu'il n'y a qu'une probabilité de 0.005 de dépasser 10.597. On peut en déduire que la probabilité de dépasser 16.828 est encore bien plus petite. La conclusion est que, partant d'une hypothèse de prévalence identique de la dysplasie dans les 3 robes, on arrive à un résultat qui est excessivement peut probable ($P << 0.005$). On peut en conclure que notre hypothèse de prévalence identique est probablement erronée, et que, par conséquent, il y a des différences de prévalences entre les robes. En regardant le tableau ci-dessus, on peut constater que la prévalence est plus élevée chez les chiens de robe chocolat, et moins élevée chez les chiens dorés.

- Solution Excel
  Une fois les effectifs observés et attendus obtenus, Excel dispose d'une fonction TEST.KHIDEUX, qui fait le reste des calculs à notre place. Si on suppose que les effectifs observés sont dans les cellules A1 à A3 (une plage

notée A1:A3 dans les formules d'Excel) et que les attendus correspondants sont dans les cellules B1 à B3, on peut taper la formule:

$$= TEST.KHIDEUX(A1 : A3; B1 : B3)$$

qui retourne la valeur 0.00022173. Cette valeur est la probabilité que nous souhaitions obtenir plus haut, c'est-à-dire la probabilité d'obtenir une valeur supérieure ou égale à 16.828 en échantillonnant une distribution de $\chi^2$ avec 2 degrés de liberté. Pour s'en convaincre, il existe une autre fonction d'Excel qui permet de passer d'une probabilité à la valeur correspondante: la fonction KHIDEUX.INVERSE, qui prend 2 arguments: le premier est la probabilité, le second le nombre de degrés de liberté. Pour notre exemple, si on tape:

$$= KHIDEUX.INVERSE(0.00022173; 2)$$

Excel nous retourne la valeur 16.8280922, qui est bien la valeur de $\chi^2$ obtenue précédemment. Alternativement, la fonction LOI.KHIDEUX($\chi^2$;dl) permet de calculer la probabilité de dépasser la valeur fournie de $\chi^2$ dans une distribution avec dl degrés de liberté. Ainsi:

$$= LOI.KHIDEUX(16.828; 2)$$

retourne une probabilité de 0.00022173, qui est celle déjà obtenue plus haut.

- Solution "R"
  R offre des possibilités très similaires à Excel. On peut calculer la probabilité associée à une valeur donnée de $\chi^2$ en utilisant la fonction:

$$pchisq(q = \chi 2, df = dl, lower.tail = FALSE)$$

Pour notre exemple:

```
> pchisq(q=16.828,df=2,lower.tail=FALSE)
[1] 0.0002217411
```

Il existe également une fonction qui effectue le test de $\chi^2$ que nous avons décrit plus haut:

$$chisq.test(x, p, correct = FALSE)$$

Le premier paramètre est le vecteur des effectifs observés. Le second paramètre est un vecteur de probabilités associées à chaque classe. Le troisième paramètre spécifie qu'il ne faut pas effectuer de correction de Yates (voir syllabus), correction qui est appliquée par défaut. La fonction calcule donc les effectifs attendus sur base des probabilités fournies et de l'effectif total des observés:

$$chisq.test(x, p = probas, correct = FALSE)$$

On peut faire le calcul avec cette fonction pour montrer qu'on obtien des résultats équivalents aux calculs précédents:

```
> chisq.test(x=c(80,48,31),p=c(0.6,0.3,0.1),correct=FALSE)
[1] 0.0002217411
```

Si on désire voir les effectifs attendus que R a calculé, on peut procéder comme suit:

```
> c<-chisq.test(x=c(80,48,31),p=c(0.6,0.3,0.1),correct=FALSE)
> names(c)
[1] "statistic" "parameter" "p.value"   "method"    "data.name"
[6] "observed" "expected"  "residuals" "stdres"
> c$expected
[1] 95.4 47.7 15.9
```

Cet exemple montre donc que la fonction chisq.test retourne en réalité un objet dont on peut demander des caractéristiques particulières, comme, par exemple, la valeur p associée au test ou les effectifs attendus.
◀

**Exercice résolu 1.3.8**

Un vétérinaire praticien pense avoir observé des effets secondaires dermatologiques lors de la prise d'un anti-parasitaire digestif. Il consulte les fiches de ses clients, et rassemble un échantillon aléatoire de celles-ci dans le tableau suivant:

| Observés | Problèmes | Indemnes | Totaux |
|---|---|---|---|
| Produit suspect | 5 | 1 | 6 |
| Autres produits | 2 | 4 | 6 |
| Totaux | 7 | 5 | 12 |

Peut-il conclure que le produit qu'il suspecte est effectivement à la base de problèmes secondaires ?
▶

Encore une fois, ce problème est un test d'hypothèse. Cette fois, l'hypothèse nulle testée est:

$$H_0 : \pi(produit) = \pi(tmoin)$$

où $\pi$ représente la prévalence des cas de problèmes dermatologiques. On reconnait la disposition sous la forme de tables de contingences, et on peut vouloir calculer les effectifs attendus de la manière habituelle:

| Attendus | Problèmes | Indemnes | Totaux |
|---|---|---|---|
| Produit suspect | $\frac{7*6}{12}$ | $\frac{5*6}{12}$ | 6 |
| Autres produits | $\frac{7*6}{12}$ | $\frac{5*6}{12}$ | 6 |
| Totaux | 7 | 5 | 12 |

On peut constater que certains effectifs attendus sont inférieurs à 5 (en fait, ils le sont tous), ce qui supprime la possibilité de tester l'hypothèse via un test de $\chi^2$. L'alternative est d'utiliser un test exact de Fisher.

– Solution "manuelle"
Si on considère le nombre "attendu" de malades ayant reçu le produit suspect si l'hypothèse nulle d'absence d'effet secondaire est vraie, soit $\frac{7*6}{12} = 3.5$, on s'aperçoit que le nombre effectivement observé de malades (5) est supérieur, ce qui pourrait soutenir la thèse du praticien. La question qu'il faut se poser est alors: est-ce qu'avoir 5 malades, ou plus, est vraisemblable s'il n'y a en fait pas d'effet secondaire (c'est-à-dire, si l'hypothèse nulle est vraie) ? On peut répondre à cette question avec un calcul de probabilités hypergéométriques. Tout d'abord, il existe une autre situation qui, avec ces effectifs, s'écarte encore plus fort de ce qui était prévu par l'hypothèse nulle:

| Observés | Problèmes | Indemnes | Totaux |
|---|---|---|---|
| Produit suspect | 6 | 0 | 6 |
| Autres produits | 1 | 5 | 6 |
| Totaux | 7 | 5 | 12 |

Représentant les deux situations par le nombre $X$ de problèmes chez les individus traités avec le produit suspect, on peut donc écrire que la probabilité de s'écarter aussi fort, voire plus fort, de ce qui était prédit (par l'hypothèse nulle) vaut:

$$P = P(X = 5) + P(X = 6)$$

Ce calcul de probabilité se fait avec la loi hypergéométrique. En effet, si on répartit les 12 animaux en deux groupes de tailles respectives 6 (comme les animaux traités avec le produit suspect) et 6 (comme les témoins), on cherche la probabilité qu'en ayant 7 animaux affectés au hasard parmi les 12, X tombe parmi le premier lot de 6 (et, forcément, (7-X) parmi le second lot de 6). Le calcul de cette probabilité est donc:

$$P = \frac{C_6^5 * C_6^2}{C_{12}^7} + \frac{C_6^6 * C_6^1}{C^7/12} = \frac{15+1}{132} = 0.121212$$

Comme cette probabilité est supérieure aux seuils de signification habituels (typiquement, $\alpha = 5\%$), on décide d'accepter l'hypothèse nulle: même si la tendance observée par le praticien va dans le sens d'un effet secondaire, on a pas démontré cet effet de manière significative.

– Solution "Excel"
Le raisonnement fait au point précédent permet de réduire les calculs avec Excel à des calculs hypergéométriques, qui peuvent être faits facilement avec la fonction:

$$LOI.HYPERGEOMETRIQUE(m_e; n_e; m_t; n_t)$$

où $m_e$ et $m_t$ représentent le nombre d'individus qui présentent la caractéristique d'intérêt (ici, le fait de présenter des effets secondaires dermatologiques) parmi l'échantillon (ici, les individus traités avec le produit incriminé: $m_e = 5$) et parmi la population (ici, le lot total: $m_t = 7$), et $n_e$ et $n_t$ sont les tailles de l'échantillon ($n_e = 6$) et de la

"population" ($n_t = 12$) respectivement. Le calcul se fait donc avec les formules:

$$= LOI.HYPERGEOMETRIQUE(5; 6; 7; 12)$$

et

$$= LOI.HYPERGEOMETRIQUE(6; 6; 7; 12)$$

qui fournissent les valeurs 0.113636 et 0.007576, dont l'addition vaut 0.121212.

– Solution "R"
R permet d'effectuer u n test exact de Fisher avec la fonction:

$$fisher.test(table, alternative = "g")$$

Le premier argument est la table de contingence, fournie à R sous la forme d'une matrice. Pour notre exemple, on peut par exemple utiliser la fonction:

$$matrix(c(5, 1, 2, 4), byrow = TRUE, nr = 2)$$

pour spécifier notre tableau de données (voir sur le site pour les explications des arguments de cette fonction). Le second argument donne l'hypothèse alternative. R raisonne sur le "odds-ratio" (cfr syllabus), et l'hypothèse nulle testée est $H_0 : OR = 1$. Les alternatives sont $H_1 : OR \neq 1$, notée dans R 'alternative="two.sided"', $H_2 : OR > 1$, notée dans R 'alternative="g"' et $H_3 : OR < 1$, notée dans R 'alternative="l"'. Dans le cas qui nous occupe, on s'intéresse à la question de savoir si l'utilisation du traitement augmente les problèmes dermatologiques, ce qui revient à tester $H_2$. Le code est donc:

```
> m<- matrix(c(5,1,2,4),byrow=T,nr=2)
> fisher.test(m,alternative="g")->ft
> ft$p.value
[1] 0.1212121
```

Comme plus haut, R retourne un objet dont on peut demander à voir les caractéristiques, comme la p-valeur affichée ici.

◄

**Exercice 1.3.9**

Recommencez l'exercice précédent mais avec tous les effectifs multipliés par 5. Comparez les solutions obtenues avec un test de Fisher et avec un $chi^2$.

►

**Solution**: la valeur p bilatérale obtenue avec le test exact de Fisher vaut 0.000181. La valeur p (toujours bilatérale avec le test de $c^2$ vaut 0.00008568.

◄

**Exercice 1.3.10**

La fréquence cardiaque au repos chez le cheval dépend elle de la race ? Pour tenter d'apporter une réponse à cette question, des chevaux appartenant à 3 races ont été sélectionnés au hasard, et leur fréquence cardiaque a été mesurée (en b/m: battements/minute). Les résultats sont reportés dans le tableau qui suit:

| Fréquence | $< 60$ b/m | $> 60$ b/m |
|---|---|---|
| Shetlands | 8 | 12 |
| Frisons | 10 | 10 |
| Purs-sangs | 6 | 14 |
| Totaux | 24 | 36 |

Que pouvez-vous conclure avec ces données ?
►

**Solution**: le test de $\chi^2$ conduit à une valeur $p = 0.4346$. Aucune différence significative n'a donc été détectée.
◄

### 1.3.4 Loi hypergéométrique

**Prérequis**

- notions de tirages avec et sans remise
- notions de probabilités conditionnelles
- formule de la loi hypergéométrique

**Exercice résolu 1.3.11**

Dans un élevage de 100 bergers allemands, 8 animaux sont testés aléatoirement pour détecter l'éventuelle présence de *Microsporum canis*, un champignon impliqué dans la plupart des cas de teigne chez le chien. Si 20 chiens de l'élevage sont porteurs, quelle est la probabilité de détecter la présence de ce pathogène dans l'élevage?
►

- Solution "Manuelle"
  Si aucun des chiens testés n'est porteur du champignon, l'exploitation sera injustement classifiée comme indemne. Dans tous les autres cas (c'est-à-dire quand il y aura entre 1 et 20 porteurs dans l'échantillon), le statut d'élevage contaminé pourra être attribué à raison. La probabilité de détection peut donc s'écrire:

$$P_d = P(r = 1) + P(r = 2) + \cdots + P(r = 20) = 1 - P(r = 0)$$

  où $r$ dénote le nombre de porteurs dans l'échantillon. Le terme à droite du symbole d'égalité est évidemment beaucoup plus facile à calculer. Le calcul peut se faire explicitement par le calcul des probabilités, ou plus implicitement via la formule de la loi hypergéométrique. Commençons

pas le calcul explicite. Si on note $s_i$ l'événement "l'individu i est sain", on peut écrire la probabilité d'avoir 0 contaminé dans notre échantillon de taille 20 de la manière suivante:

$$P = P(s_1) * P(s_2|s_1) * P(s_3|S_1, s_2) * \cdots * P(s_{20}|s_1, \cdots, s_{19})$$

Dans cette expression:

$$\begin{aligned} P(s_1) &= \frac{80}{100} \\ P(s_2|s_1) &= \frac{79}{99} \\ P(s_3|s_1, s_2) &= \frac{78}{98} \\ \cdots &= \cdots \\ P(s_{20}|s_1, s_2, \cdots, s_{19}) &= \frac{61}{81} \end{aligned}$$

Le calcul conduit donc à:

$$P = \frac{80 * 79 * 78 * \cdots * 61}{100 * 99 * 98 * \cdots * 81} = \frac{80!/60!}{100!/80!} = \frac{80! * 80!}{100! * 60!} = 0.0066$$

On peut calculer cette expression en faisant le produit des fractions données plus haut. Alternativement, on peut utiliser la formule de la loi hypergéométrique, qui est d'application dans ce cas où on regarde un événement binaire et où la probabilité à chaque tirage varie (tirage "sans remise"). La probabilité d'avoir 0 des 20 contaminés (et, par conséquent, 20 parmi les non contaminés) quand on échantillonne 20 individus au hasard parmi les 100 se calcule par:

$$P = \frac{C_{20}^0 * C_{80}^{20}}{C_{100}^{20}} = \frac{80!}{20! * 60!} * \frac{80!}{20! * 60!} * \frac{80! * 20!}{100!} = \frac{80! * 80!}{100! * 60!}$$

Cette probabilité est bien identique à celle découverte plus haut. On peut en déduire que la probabilité de détection est très élevée:

$$P_d = 1 - 0.0066 = 0.9934$$

- Solution "Excel"
  Avec "Excel", on peut faire le calcul explicite facilement:
  - On écrit 100 en A1 et 80 en B1,
  - On écrit la formule "=A1-1" en A2
  - On "copie" la cellule A2, et on la colle sur la plage A2:B20,
  - On écrit la formule "=B1/A1" en C1,
  - On "copie" la cellule C1, et on la colle sur la plage C2:C20,
  - On multiplie les probabilités en écrivant en D1 la formule "=PRODUIT(C1:C20)".

Le résultat (0.0066) est alors soustrait de 1 pour obtenir le résultat final.
De manière plus simple, on peut aussi utiliser la fonction LOI.HYPERGEOMETRIQUE en tapant:

$$= 1 - LOI.HYPERGEOMETRIQUE(0; 20; 20; 100)$$

qui fournit le même résultat.

- Solution "R"
Les deux calculs (explicite et implicite) peuvent se faire facilement en utilisant les fonctionnalités de R. Le listing qui suit montre les deux calculs, avec, pour le second, l'utilisation de la fonction

$$phyper(a, b, c, d, lower.tail = TRUE)$$

Dans cette expression, a est le nombre de fois que l'événement qui nous intéresse a lieu (ici, l'événement est 'tirer un individu contaminé', et par conséquent, $a = 0$), b est le nombre de fois où l'événement se produit dans la 'population' (ici, $b = 20$), c est le nombre de fois où le contraire de l'événement se produit dans la population (ici, $c = 80$) et d est le nombre de tirages (ici, $d = 20$). Comme d'habitude, le dernier paramètre permet de préciser si on souhaite $P(r \leq a)$ (lower.tail=TRUE) ou $P(r > a)$ (lower.tail=FALSE).

```
> numer<-61:80
> denom<-81:100
> prod(numer/denom)
[1] 0.006595944
> phyper(0,20,80,20)
[1] 0.006595944
```

◀

**Exercice 1.3.12**

Dans le problème précédent, quelle devrait être la taille minimale de l'échantillon pour que la probabilité de détection soit supérieure à 80% ?

▶

**Solution**: la taille minimale est $n = 7$. ◀

### 1.3.5 Loi normale

**Prérequis**

- calcul de probabilités et distributions continues
- théorème de la limite centrale

**Exercice résolu 1.3.13**

Si le poids des bovins d'une race à viande est considéré comme distribué normalement, avec une moyenne de 700 kilos et une déviation standard de 100 kilos, quelle est la probabilité qu'un échantillon de 10 bovins de cette race ait un poids moyen de plus de 750 kilos ?

▶

- Solution "manuelle"
  Les paramètres importants, quand on travaille avec une distribution normale, sont la moyenne et la déviation standard de cette distribution. Dans l'exercice, la moyenne est $\mu_P = 750$ kilos, et la déviation standard est $\sigma_P = 100$ kg. On pourrait, avec la distribution correspondante, notée $N(700; 100)$, calculer les probabilités associées à des poids individuels. Toutefois, on parle ici de poids moyen d'un groupe de 10 bovins et non pas de poids individuels. La conséquence est triple (théorème de la limite centrale):

  - La distribution des poids moyens tend vers une distribution normale,
  - La moyenne de la distribution des poids moyens est la même que celle des poids individuels (soit, $\mu_{\bar{P}}$=700 kg),
  - La déviation standard de la distribution des moyennes est plus petite que celle de la distribution des poids individuels, le facteur de proportionnalité étant la racine de la taille de l'échantillon (soit $\sigma_{\bar{P}} = \sigma_P/\sqrt{10}$ kg)

  On peut dès lors facilement calculer la valeur de z associée à ce poids moyen de 750 kilos:

  $$z = \frac{\bar{P} - \mu_{\bar{P}}}{\sigma_{\bar{P}}} = \frac{750 - 700}{100/\sqrt{10}} = 0.5 * \sqrt{10} = 1.58$$

  On peut utiliser cette valeur de z et une table de z comme celle du syllabus pour en déduire:

  $$P(\bar{P} > 750) = P(z > 1.58) = 0.5 - 0.4429 = 0.0571$$

- Solution "Excel"
  Le même calcul peut se faire directement avec Excel (c'est-à-dire, sans passer par z). Il faut pour cela utiliser la fonction LOI.NORMALE de la manière suivante:

  $$= 1 - LOI.NORMALE(750; 700; 100/RACINE(10); VRAI)$$

  Dans cette fonction, le premier paramètre est la valeur pour laquelle on recherche une probabilité, le second est la moyenne de la distribution, le troisième est la déviation standard de la distribution, et le dernier est un paramètre qui précise si on veut calculer une densité de probabilité (FAUX) ou une probabilité. Comme la fonction retourne la probabilité que le poids moyen soit inférieur à 750, on obtient la probabilité recherchée en utilisant le complément. Le résultat est $P = 0.056923$.

- Solution "R"
  On utilise la fonction:

  $$pnorm(x, mean = m, sd = s, lower.tail = TRUE)$$

  où x est la valeur pour laquelle on calcule la probabilité, m est la moyenne de la distribution utilisée, s est la déviation standard utilisée et lower.tail

indique si on souhaite $(P \leq x)$ ou $(P > x)$ (par défaut, 'lower.tail=TRUE', ce qui signifie qu'on souhaite obtenir $P \leq x$). Pour effectuer le calcul qui nous intéresse, il faut donc utiliser:

```
> pnorm(750,700,100/sqrt(10),lower.tail=FALSE)
[1] 0.05692315
```

◀

**Exercice 1.3.14**

Dans toutes les races de moutons marocaines, le poids des agneaux à la naissance est supposé être optimale entre 3 et 4.5 kilos. Deux éleveurs comparent leurs performances. Le premier, éleveur de D'man, où le poids à la naissance suit une distribution normale de moyenne 3 kilos et de déviation standard 0.5 kilos a obtenu 30 agneaux, avec une moyenne de poids de 3.2 kilos. Le second, qui travaille sur la race Timahdit, de poids moyen à la naissance distribué normalement avec une moyenne égale à 3.5 kilos et une déviation standard égale à 0.6 kilos a obtenu une moyenne de 3.8 kilos pour les 20 agneaux qu'il a obtenus cette année. Tous les deux prétendent avoir obtenu des performances exceptionnelles en ce qui concerne le poids à la naissance. Lequel a obtenu les performances les plus exceptionnelles ? Le terme "exceptionnel" vous semble-t-il justifié ?

▶

**Solution**: la probabilité d'avoir de si bons résultats, voire encore meilleurs, vaut chez la D'man $p = 0.01423$ et chez la Timahdit $p = 0.01267$. Les performances de la Timahdit sont donc légèrement meilleures encore que celles de la D'man, mais les résultats dans les deux races sont effectivement exceptionnels puisqu'ils se situent tous les 2 dans les 2 derniers percentiles de la distribution des poids à la naissance.

◀

## 1.4 Échantillonnage (Séances TP 3 et 5)

Dans la plupart des problèmes d'analyse de données, les conclusions qui seront tirées reposent sur les échantillons de données récoltées: on essaie d'inférer de l'information sur les populations visées en utilisant l'image imparfaite dont on dispose à travers un échantillon. Le processus d'échantillonnage est donc central. Pour nous permettre de disposer d'échantillons correspondant aux situations étudiées, Excel et R disposent de nombreuses fonctionnalités qui permettent de simuler un échantillonnage réel. Nous allons utiliser ces fonctionnalités dans différents contextes dans ce chapitre.

### 1.4.1 Échantillonnage en loi uniforme (discrète et continue)

Pour commencer avec un problème simple illustrant l'approche, nous allons nous intéresser au problème suivant: comment simuler un jet de dés (à 6 faces) à l'aide des logiciels ? La solution est simple. En Excel, on peut utiliser la fonction $ALEA()$ vue plus haut, ou la fonction $ALEA.ENTRE.BORNES(< min >; < max >)$. La fonction $ALEA()$, vue plus haut, permet de générer un

nombre réel entre 0 et 1. Multipliant le nombre obtenu par 6, on obtient un nombre réel entre 0 et 6. Si on prend la partie entière de ce nombre réel, on obtient 0, 1, 2, 3, 4 ou 5, chaque chiffre ayant la même probabilité puisque la distribution est uniforme (ce qui signifie que, si on appelle x la valeur fournie par ALEA():

$$P(\frac{0}{6} < x < \frac{1}{6}) = P(\frac{1}{6} < x < \frac{2}{6}) = P(\frac{2}{6} < x < \frac{3}{6}) = P(\frac{3}{6} < x < \frac{4}{6}) = P(\frac{4}{6} < x < \frac{5}{6}) = P(\frac{5}{6} < x < \frac{6}{6})$$

ce qui conduit à des nombres entiers équiprobables). Il ne reste plus qu'à ajouter 1 au résultat pour avoir les 6 valeurs faciales du dé avec une même probabilité. La formule est donc:

$$= ENT(6 * ALEA()) + 1$$

où $ENT(< nombre >$ retourne la partie entière de $< nombre >$. La fonction $ALEA.ENTRE.BORNES(< min >; < max >)$ permet d'effectuer cette opération de manière encore plus directe, puisqu'elle retourne un nombre entier compris entre les limites (incluses) définies par $< min >$ et $< max >$. On simule donc un dé en tapant:

$$= ALEA.ENTRE.BORNES(1; 6)$$

Une fonction similaire existe en R: la fonction $runif(n, min =< min >, max =< max >)$, dans laquelle $n$ est le nombre de valeurs à générer (par exemple, le nombre de jets de dé), $< min >$ est la limite inférieure (0, par défaut) et $< max >$ est la limite supérieure (1, par défaut). On peut donc simuler un dé en utilisant les 2 méthode vues pour Excel:

```
> # Methode 1 de simulation d'un de (10 jets)
> floor(6*runif(10))+1
[1] 1 5 1 1 4 3 3 5 5 1
> # Methode 2 de simulation d'un de (10 jets)
> floor(runif(10,1,7))
[1] 4 4 6 1 6 5 1 6 3 1
```

**Exercice résolu 1.4.1**

Obtenez de manière théorique et de manière empirique (par simulation) la probabilité que la somme de deux dés soit supérieure à 9.

►

Théoriquement, il y a 36 combinaisons possibles pour les 2 dés. Parmi ces combinaisons, 6 correspondent à des sommes supérieures à 9: 4-6, 5-5, 5-6, 6-4, 6-5, 6-6. La probabilité est donc P = 6/36 = 1/6 = 0.1667.
Empiriquement, on pourrait simuler un grand nombre de jets de 2 dés et comptabiliser le nombre de jets donnant des valeurs supérieures à 9. La procédure en R est la suivante:

```
> # Effectuons 10000 jets
> n<-10000
> jets<-floor(runif(n,1,7))+floor(runif(n,1,7))
> # Extraction des jets dont la somme excede 9
> j9<-jets[jets>9]
> # Proportion
> length(j9)/length(jets)
[1] 0.1642
```

On voit que la solution empirique est proche de la solution théorique.

◀

**Exercice 1.4.2**

Calculez de manière théorique et empirique la probabilité d'obtenir une suite (3 valeurs qui se suivent) quand on jette 3 dés.

▶

**Solution**: la probabilité théorique est P = 1/9 = 0.111.

◀

Une des situations où l'échantillonnage dans des distributions uniforme est intéressant dans des applications plus vétérinaires est quand on compare des effectifs, comme dans l'exemple qui suit.

**Exercice résolu 1.4.3**

Un traitement par vaporisation d'une solution fongicide est supposé protéger les ruches contre une attaque par un champignon potentiellement létal pour les abeilles. Pour tester ce traitement, 2 cohortes de 100 abeilles sont constituées aléatoirement. Une des cohortes subit la vaporisation alors que l'autre pas. Les deux cohortes sont ensuite exposées aux champignons, et les mortalités sont relevées dans les deux groupes. Simulez cette expérience en supposant que les taux de survie dans la population (pas nécessairement dans l'échantillon...!) chez les abeilles ayant subi la vaporisation est de 80% alors qu'il est de 75% chez celles qui n'ont pas été vaporisées. Remarquez qu'en pratique, ces taux de survie sont évidemment inconnus: si on savait dès le départ que le taux de survie est plus élevé en vaporisant, il n'y aurait plus de raison de faire l'expérience ! Dans quelle proportion de cas pourra-t-on démontrer la (relative) efficacité du traitement fongicide ?

▶

Nous allons travailler avec R, mais une solution avec Excel peut être obtenue de manière similaire. La première étape consiste à réaliser que, avec une distribution uniforme entre 0 et 1:

$$P(0 < x < t) = t$$

Cette propriété est assez évidente: comme la distribution a une forme rectangulaire, la surface à gauche de t est un rectangle de base t et de hauteur égale à 1. Par conséquent, pour simuler un événement qui a une probabilité $t$ de se produire, on tire un nombre aléatoirement dans une distribution uniforme entre 0 et 1. Si le nombre aléatoire est inférieur à $t$ (ce qui a donc une probabilité $P = t$ de se produire), on considère que l'événement d'intérêt s'est produit. Dans le cas contraire, on considère que l'événement ne s'est pas produit. On peut donc obtenir un échantillon aléatoire de 100 abeilles en répétant 100 fois cette procédure. Le code R est le suivant:

```
> # Echantillon d'abeilles non vaporisees
> cohort_v<-runif(100)
> survit_v<-length(cohort_v[cohort_v<0.8])
> mortes_v<-100-survit_v
> # Echantillon d'abeilles vaporisees
> cohort_nv<-runif(100)
> survit_nv<-length(cohort_nv[cohort_nv<0.75])
> mortes_nv<-100-survit_nv
```

A ce stade, nous avons obtenu les deux échantillons aléatoires, il nous reste à savoir si la différence entre les deux est significative. Le test de $\chi^2$ permet d'effectuer ce test:

```
> # Table de contingences et test de khi deux
> table<-matrix(c(survit_v,mortes_v,survit_nv,mortes_nv),nr=2)
> khi2<-chisq.test(table,correct=F)
> khi2

        Pearson's Chi-squared test

data:  table
X-squared = 0.027, df = 1, p-value = 0.8694
> khi2$p.value
[1] 0.8694083
```

On voit que pour cet exemple particulier, le test ne voit pas de différence significative entre les deux cohortes. Pour obtenir la probabilité d'observer une différence significative (qui est bien réelle, puisque les données ont été générées en tenant compte de cette différence), il faudrait répéter un grand nombre de fois cette expérience, et comptabiliser la proportion de situations où une différence significative (au seuil $\alpha = 5\%$, par exemple) est détectée. Le code suivant, qui intègre les étapes précédentes, effectue ces opérations:

```
> # Compteur de situations significatives
> n_sig<-0
> # Boucle: on repete 1000 fois l'experience
> for (i in 1:1000) {
+  # Echantillon d'abeilles vaporisees
+  cohort_v<-runif(100)
+  survit_v<-length(cohort_v[cohort_v<0.8])
+  mortes_v<-100-survit_v
+  # Echantillon d'abeilles non vaporisees
+  cohort_nv<-runif(100)
+  survit_nv<-length(cohort_nv[cohort_nv<0.75])
+  mortes_nv<-100-survit_nv
+  # Table de contingence
+  table<-matrix(c(survit_v,mortes_v,survit_nv,mortes_nv),nr=2)
+  # Test
+  khi2<-chisq.test(table,correct=F)
+  if (khi2$p.value<0.05) { n_sig<-n_sig+1 }
+ }
> # Proportion de situations significatives
> n_sig/1000
[1] 0.132
```

On constate donc que, bien qu'il y ait une différence entre les 2 populations (80% de survie dans l'une et 75% de survie dans l'autre), le test avec deux cohortes de 100 individus ne permet de ne constater cette différence que dans une proportion modeste des cas (13.2% des cas dans cette simulation, vos chiffres pourraient être légèrement différents).

◀

**Exercice 1.4.4**

Une expérience transversale ("cross-sectional") est effectuée pour vérifier si les juments et les étalons d'une race de cheval de sport ont la même sensibilité à un problème de tendinite au boulet. La réalité, qu'en pratique on ne connait pas, est que 10% des femelles et 15% des mâles sont concernés par ces problèmes. On supposera en outre qu'il y a 48% de mâles et 52% de femelles dans cette population. Recherchez, comme dans le problème précédent, la proportion de

situations dans lesquelles la différence entre mâles et femelles sur ce point pourra être mise en évidence dans une expérience qui concerne 300 chevaux.

►

**Solution**: la proportion obtenue dans un ensemble de simulations est P = 0.263. Vos solutions pourraient être légèrement différentes.

◄

### 1.4.2 Échantillonnage en loi normale

De nombreux caractères ont une distribution normale (le poids, la taille ainsi que de nombreux paramètres physiologiques, etc...). Si on veut simuler une expérience menée dans ce contexte, l'échantillonnage dans une distribution uniforme ne convient donc plus: il faut cette fois imiter l'échantillonnnage dans une distribution normale. Le principe général est similaire: on tire une surface (c'est-à-dire une probabilité) au hasard entre 0 et 1, et on déduit la valeur de la variable correspondante. Le schéma 1.15 donne l'idée de la manière de procéder. Aussi bien Excel que R permette de faire ce genre de manipulations. En Excel, on peut utiliser:

$$= LOI.NORMALE.INVERSE(< surface >; < moyenne >; < dviation_standard >)$$

Par exemple, si on raisonne sur une variable normale X qui a une moyenne de 100 et une déviation standard de 10, l'expression:

$$= LOI.NORMALE.INVERSE(0.6; 100; 10)$$

retourne 102.523471, c'est-à-dire la valeur de X qui est telle que 60% des observations lui sont inférieures. Si on souhaite tirer une valeur au hasard dans cette distribution, il suffit de prendre une surface au hasard (avec la fonction ALEA). Ainsi,

$$= LOI.NORMALE.INVERSE(ALEA(); 100; 10)$$

retourne une valeur aléatoire de X, tirée dans la distribution qui nous intéresse. Le principe est similaire avec R, où la fonction à utiliser est:

$$rnorm(n, mean = \mu, sd = \sigma)$$

où n désigne le nombre de valeurs aléatoires à générer, et $\mu$ et $\sigma$ sont les moyenne et déviation standard de la distribution normale visée. L'exemple qui suit illustre cette fonction d'échantillonnage.

**Exercice résolu 1.4.5**

Illustrez le "théorème de la limite centrale" en montrant que si on échantillonne un grand nombre de fois une distribution de moyenne $\mu$ et de déviation standard $\sigma$, les moyennes des échantillons obtenus tendent à se distribuer de manière gaussienne avec une moyenne valant $\mu$ et une déviation standard valant $\sigma/\sqrt{n}$ où $n$ est la taille de l'échantillon.

►

Nous allons illustrer ce résultat 2 fois: une première fois en partant d'une distribution uniforme, puis une seconde en partant d'une distribution normale.
Commençons par la distribution uniforme. Si on suppose que X peut prendre des valeurs entre 0 et 6 (par exemple), alors la densité prend une forme de

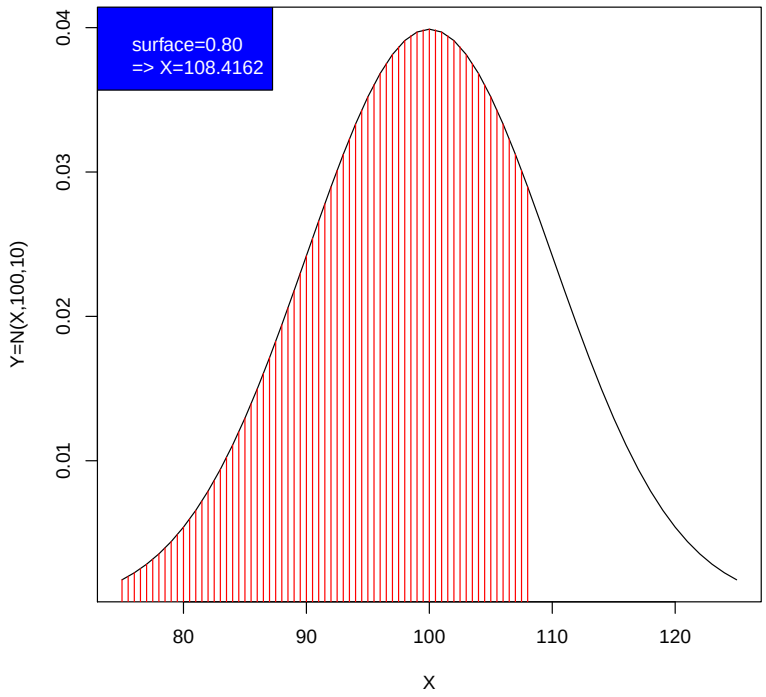



Figure 1.15: Échantillonnage dans une distribution normale: la surface, prise au hasard entre 0 et 1, correspond à une valeur unique de la variable distribuée normalement X. La distribution utilisée dans le schéma est une normale de moyenne $\mu = 100$ et de déviation standard $\sigma = 10$.

rectangle de largeur 6 et de hauteur 1/6. Mathématiquement, on peut écrire: $f(x) = 1/6$ pour X dans l'intervalle $[0; 6]$ et $f(x) = 0$ ailleurs. On peut calculer les paramètres de cette distribution:

$$\begin{aligned}
\mu &= \int_{X=-\infty}^{X=+\infty} X * f(X)dx \\
&= \int_{X=0}^{X=+6} X * (1/6)dx \\
&= (1/6) * \left[x^2/2\right]_0^6 \\
&= 18/6 \\
&= 3
\end{aligned}$$

$$\begin{aligned}
\sigma^2 &= \int_{X=-\infty}^{X=+\infty} (X-\mu)^2 * f(X)dx \\
&= \int_{X=0}^{X=+6} (X-3)^2 * (1/6)dx \\
&= (1/6) * \left[(x-3)^3/3\right]_0^6 \\
&= 1/6 * [(27+27)/3] \\
&= 3
\end{aligned}$$

Prélevons à présent 1000 échantillons de taille 9 (par exemple) dans cette distribution, et calculons les moyennes $\bar{X}_i$ pour chaque échantillon. Une fois les 1000 moyennes calculées, nous calculerons la moyenne des moyennes $m_{\bar{X}}$ et la variance des moyennes $s^2_{\bar{X}}$. Le code R est le suivant:

```
> # Creer un vecteur pour contenir les 1000 moyennes
> means<-vector(mode="numeric",length=1000)
> # Generer 1000 echantillons
> for (i in 1:1000) {
>  echant<-runif(9,min=0,max=6)
>  means[i]<-mean(echant)
> }
> # Calcul de la moyenne et de la variance
> mean(means)
[1] 3.010592
> var(means)
[1] 0.3358332
```

On voit que les résultats de cette simulation sont très proches des résultats théoriques $\mu_{\bar{X}} = \mu_X = 3$ et $\sigma^2_{\bar{X}} = \sigma^2_X/9 = 1/3$. Si on recommence en échantillonnant une distribution normale de moyenne et de variance valant 3 toutes les deux, seule l'instruction de génération de l'échantillon change:

```
> # Creer un vecteur pour contenir les 1000 moyennes
> means<-vector(mode="numeric",length=1000)
> # Generer 1000 echantillons
> for (i in 1:1000) {
>  echant<-rnorm(9,mean=3,sd=sqrt(3))
>  means[i]<-mean(echant)
> }
> # Calcul de la moyenne et de la variance
> mean(means)
[1] 2.991767
```

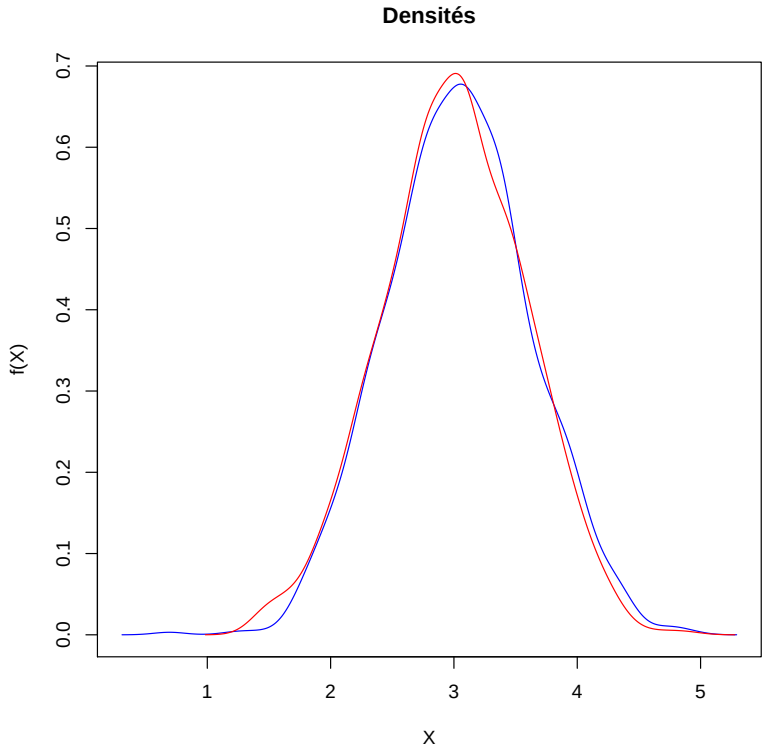



Figure 1.16: Densités de probabilités associées à l'échantillonnage dans une distribution uniforme (en rouge) et dans une distribution normale (en bleu): les échantillons, dans les 2 cas, proviennent de distributions de moyenne et de variance valant 3 et les tailles d'échantillon valent 9.

```
> var(means)
[1] 0.3459316
```

A nouveau, les valeurs sont extrêmement proches de ce qui était attendu. On peut aussi vouloir représenter graphiquement la densité obtenue à partir de ces 1000 échantillons. Le résultat est montré dans la figure 1.16: dans les deux simulations, la densité se rapproche d'une densité normale.

◀

**Exercice 1.4.6**

Si on échantillonne dans une distribution normale de moyenne valant $\mu$ et de déviation standard valant $\sigma$, et qu'on calcule la moyenne $\bar{X}$ et la déviation standard $s$ de l'échantillon, on peut calculer la grandeur $t = (\bar{X} - \mu)/(s/\sqrt{n}$ où $n$ est la taille de l'échantillon. Montrez que, si $n$ n'est pas trop grand (typiquement, inférieur à 10), la distribution obtenue pour $t$ en répétant l'expérience prend une forme de distribution, mais n'est pas une distribution normale (quel type de distribution obtient on ?). Que se passe-t-il si $n$ augmente ?

►

**Solution**: on peut effectuer la comparaison entre la distribution obtenue et la distribution normale standard ($\mu = 0$ et $\sigma = 1$) de manière graphique, en représentant les densités correspondant aux 2 distributions. On emploie alors:

$$plot(x, y, type =' l', col =' red')$$

pour représenter graphiquement la première densité (celle de t), où $x$ et $y$ sont deux vecteurs des abscisses et ordonnées des points de la première distribution. Ces coordonnées sont fournies via:

$$dens1 < -density(t)$$

$dens1$ est alors un objet, dont les éléments x ($dens\$x$) et y ($dens\$y$) sont les coordonnées recherchées. ◄

### 1.4.3 Généralisation de l'échantillonnage aux autres lois

L'échantillonnage n'est évidemment pas restreint aux cas de la distribution uniforme et de la distribution binomiale. L'exemple qui suit illustre l'échantillonnage dans une distribution binomiale.

**Exercice résolu 1.4.7**

Un variant génétique du génome du poulet pourrait avoir un effet positif sur la qualité de la viande. Pour tester cette hypothèse, un généticien souhaiterait créer un élevage dans lequel les individus reproducteurs (coqs et poules) seraient porteurs du génotype favorable. Sachant que ce génotype a une fréquence de 0.3 dans la population, si ils examinent 10 élevages de 30 poulets, quelle est la probabilité qu'ils obtiennent plus de 100 individus de génotype favorable ?

►

On peut raisonner théoriquement ou empiriquement. Commençons par l'approche théorique. Si le pourcentage de variants variait entre exploitations (par exemple parce qu'il y a des races différentes dans les différentes exploitations et que la fréquence du variant est différente de race en race), il faudrait:

- calculer $P(r_i)$ pour chaque valeur de $r_i = 0, 1, \cdots, 30$ et pour chaque exploitation i où $i = 1, 2, \cdots, 10$,
- pour chaque combinaison d'effectifs $r_i$, on aurait donc un effectif $n = \sum_i r_i$, et une probabilité associée à cette combinaison qui serait $P(r_1, r_2, \cdots, r_{10}) = P(r_1) * P * r_2) * \cdots * P(r_{10})$, où chaque probabilité $P(r_i)$ peut être calculée à l'aide de la distribution binomiale,
- la probabilité recherchée s'obtiendrait alors en sommant les probabilités de toutes les situations où $n \geq 100$.

La difficulté vient évidemment du fait que le nombre de possibilités pour le calcul du second point est gigantesque: il vaut $31^{10} = 819628286980801...$ Cette approche est donc difficilement praticable. Heureusement, notre problème est plus simple puisque les probabilités sont du variant sont les mêmes dans chaque exploitation: tout se passe en fait comme si on échantillonnait dans une grande

exploitation de 300 individus, avec une probabilité d'obtenir le variant recherché valant $p = 0.3$. On a alors un "simple" problème binomial, consistant à calculer:

$$P(r \geq 100 | n = 300, p = 0.3)$$

On peut effectuer ce calcul facilement en R en tapant:

$$pbinom(99, size = 300, prob = 0.3, lower.tail = F)$$

Les deuxième et troisième paramètres définissent la distribution binomiale, le premier paramètre est la valeur de r et le quatrième précise si on veut calculer $P(\leq r)$ (dans ce cas, on utilise $lower.tail = TRUE$, qui est l'option par défaut) ou $P(> r)$ (alors, on utilise $lower.tail = FALSE$). De manière équivalente, on aurait donc pu calculer:

$$1 - pbinom(99, size = 300, prob = 0.3)$$

ou, en utilisant Excel:

$$1 - LOI.BINOMIALE(99; 300; 0.3; VRAI)$$

Dans cette formule, le dernier paramètre signifie qu'on additionne les probabilités pour toutes les valeurs inférieures ou égales à 99. Le résultat du calcul est $P = 0.1163165$.
L'approche empirique consiste à simuler l'échantillonnage un grand nombre de fois et à voir dans quelle proportions de situations on obtient au moins 100 individus présentant le variant d'intérêt. La simulation peut être faite facilement en R:

```
> # Nombre de situations ou n est superieur ou egal a 100
> n<-0
> # Boucle sur les simulations
> for (i in 1:1000) {
+   if (sum(rbinom(10,prob=0.3,size=30))>=100) { n<-n+1 }
> }
> n/1000
[1] 0.118
```

On voit que le résultat empirique est une bonne approximation du résultat théorique. Il faut remarquer la manière dont l'instruction dans la boucle échantillonne 10 fois la distribution binomiale, crée un vecteur avec ces 10 valeurs de r, les additionne, compare le résultat à 100, et incrémente le compteur si nécessaire, le tout en une seule ligne de code !

**Exercice 1.4.8**

Un responsable de la chaine alimentaire consulte les statistiques de découvertes de carcasses positives pour la présence de résidus (dioxine). Il constate qu'on découvre en moyenne 3 carcasses par mois dans chaque province (il y a 10 provinces). Quelle est la distribution du nombre de carcasses contaminées en Belgique par an ?
*Truc*: utiliser une loi de Poisson (voir la fonction rpois) et approcher le problème comme dans l'exemple précédent.
▶
**Solution**: nous donnerons, à titre indicatif, la figure représentant la distribution empirique obtenue par simulation avec 10000 échantillons

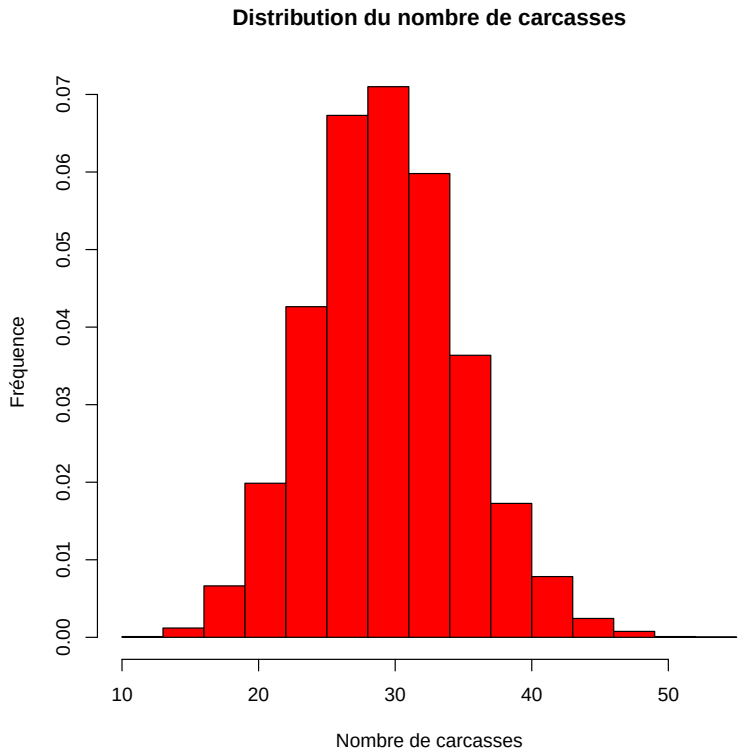



Figure 1.17: Fréquences absolues obtenues en simulant 10000 échantillons tirés dans la distribution du nombre de carcasses contaminées dans 10 provinces.

◀