Loi des grands nombres et TCL

Eric Marcon

12 septembre 2018

Résumé

L'objectif du cours est de connaître le gain de précision d'une mesure en fonction de l'effort d'échantillonnage.

Table des matières

1 Contexte **1**

2 Loi des grands nombres **2**

3 Théorème de la limite centrale **3**

4 Dimensionnement **7**

5 Conclusion **9**

1 Contexte

On mesure une variable quelconque (par exemple la concentration d'une molécule intéressante dans un fruit) et on a la possibilité de répéter mes mesures.

L'échantillonnage est correct :

- On dispose de plusieurs échantillons indépendants les uns des autres (les fruits proviennent d'arbres différents) ;
- Les échantillons représentent aussi bien que possible la population totale (on dispose de fruits de plusieurs provenances, ayant poussé dans toutes les conditions réelles, et les effectifs échantillonnés sont à peu près proportionnels à ceux de la population étudiée).

Chaque mesure est considérée comme une variable aléatoire. La valeur observée est la somme de la "vraie valeur", c'est-à-dire l'espérance de la variable aléatoire, et d'une erreur de mesure. L'erreur est aléatoire, elle provient de la variabilité individuelle (chaque fruit est un peu différent des autres), des conditions de récolte, de mesure, de l'opérateur, des réglages de l'appareil, de sa précision, etc.

En situation réelle, on ne sait rien sur la loi de la variable aléatoire qui donne la mesure.

On prend toutes les précautions pour éviter une erreur systématique : si les fruits sont systématiquement récoltés avant maturité, il est probable que la molécule recherchée soit systématiquement moins présente (ou plus présente selon les cas) ; idem si on ne récolte que des fruits d'arbres de petite taille parce qu'ils sont plus faciles à attraper ; enfin, si notre machine est mal réglée et surestime systématiquement les mesures, le résultat sera biaisé.

On ne peut pas éviter les erreurs aléatoires, qui surestiment ou sous-estiment la vraie valeur, mais se compensent : certains fruits seront plus ou moins concentrés, et l'instrument de mesure arrondit les résultats...

Le cours traite un cas théorique en simulant des mesures. Nous mesurons une quantité dont la vraie valeur est fixée à 0,5. L'opérateur n'est pas très doué et obtient avec une probabilité égale des valeurs entre 0 et 1. Dans la réalité, on obtient plus probablement des résultats proches de la vraie valeur, nous traitons donc un cas particulièrement défavorable.

2 Loi des grands nombres

La loi des grands nombres [1] dit que la répétition de la même mesure va nous permettre d'approcher la distribution théorique. Nous tirons dans une loi uniforme entre 0 et 1 pour simuler la mesure. Plus le nombre de tirages est grand, plus l'histogramme des résultats se rapproche de celui de la loi uniforme (Figure 1).

```
fh <- function(n) {
    Tirages <- runif(n)
    thePlot <- ggplot(data.frame(Tirages), aes(Tirages)) +
        geom_histogram(bins = nclass.Sturges(Tirages),
            color = "black", fill = "white", boundary = 0) +
        labs(x = "", y = "", title = paste(n, "tirages"))
    return(thePlot)
}
library("gridExtra")
grid.arrange(fh(10), fh(1000), fh(1e+05), ncol = 3,
    bottom = "Tirages", left = "Fréquence")
```

La loi uniforme[2] : est bien connue : son espérance est le centre de son intervalle, c'est-à-dire 0,5, et son écart-type est égal à son intervalle divisé par $\sqrt{12}$), c'est-à-dire environ 0,289 dans notre cas.

Dans la réalité, on va répéter les mesures et en faire la moyenne. La loi des grands nombres dit que cette moyenne se rapprochera de la vraie valeur en répétant les mesures (Figure 2).

```
NbTirages <- 1000
Tirages <- NbTirages %>% runif
Moyenne <- Tirages %>% cumsum %>% `/`(1:NbTirages)
```

[1] http ://fr.wikipedia.org/wiki/Loi_des_grands_nombres
[2] http ://fr.wikipedia.org/wiki/Loi_uniforme_continue

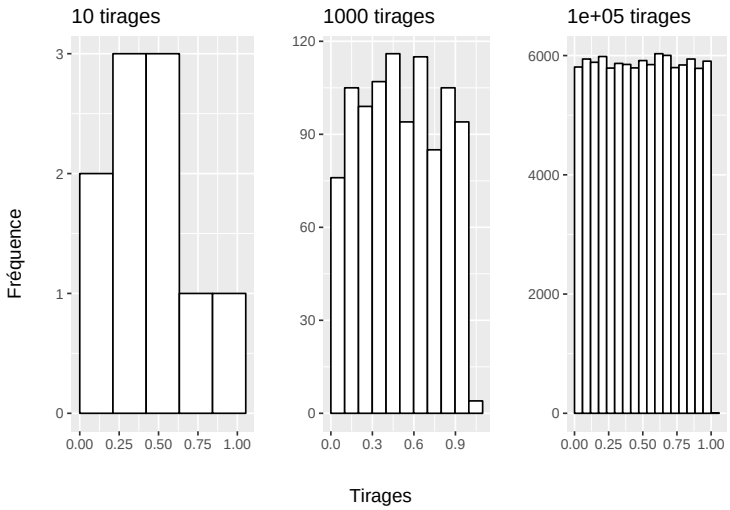

FIGURE 1: Histogramme des résultats du tirage dans une loi uniforme entre 0 et 1 pour 10, 1000 et 100000 tirages. Plus le nombre de tirages est élevé, plus l'histogramme est plat.

```
ggplot(data.frame(Tirages = 1:NbTirages, Moyenne),
    aes(x = Tirages, y = Moyenne)) + geom_line() +
    geom_hline(yintercept = 0.5, col = "red")
```

La répétition des mesures permet donc de s'approcher en moyenne de la vraie valeur. On peut vérifier que l'écart-type des mesures s'approche aussi de l'écart-type théorique :

```
sd(Tirages)
```

```
## [1] 0.2891708
```

3 Théorème de la limite centrale

Le théorème de la limite centrale[3] : dit que la distribution de la moyenne des tirages converge vers une loi normale.

Dans la réalité, on ne fera qu'une seule fois les 1000 mesures du paragraphe précédent. La question est la précision du résultat obtenu (qu'on appellera "mesure moyenne") : il sera proche de la vraie valeur, d'autant plus qu'on aura répété les mesures, mais à quel point ?

La mesure moyenne suit approximativement une loi normale (d'autant mieux que le nombre de mesures a été grand), dont l'espérance est celle de la mesure individuelle (0,5 dans notre cas) et l'écart-type celui de la mesure individuelle

[3] http ://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_central_limite

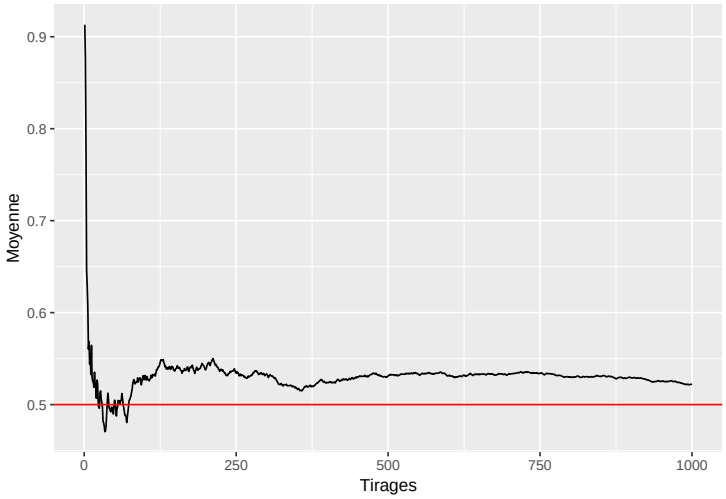

FIGURE 2: Évolution de la moyenne des résultats du tirage dans une loi uniforme entre 0 et 1 en fonction du nombre de tirages. La moyenne se rapproche de 0.5, sa valeur théorique, quand on augmente le nombre de tirages.

($1/\sqrt{12}$ dans notre cas) divisé par la racine carrée du nombre de mesures ($\sqrt{1000}$ dans notre cas). La distribution d'une loi normale est bien connue : on sait que 95% de ses réalisations sont autour de son espérance, à plus ou moins deux écarts-types. Dans notre cas, la mesure moyenne a 95% de chances de se trouver dans l'intervalle $[0,5 - 2 \times (1/\sqrt{12})/\sqrt{1000}; 0,5 + 2 \times (1/\sqrt{12})/\sqrt{1000}]$ c'est-à-dire $[0,482; 0,518]$. C'est la définition de la précision du résultat : son intervalle de confiance à 95% est entre 0,482 et 0,518. Dans la réalité, l'écart-type de la mesure individuelle est en général inconnu au début de l'expérience, on pourra seulement l'estimer à partir des données. On peut essayer d'estimer l'écart-type à partir d'un petit nombre de mesures pour avoir une idée de la variabilité : cette estimation ne sera pas très bonne parce que la loi des grands nombres ne s'applique que pour les grands nombres, mais une valeur approximative est meilleure que pas de valeur du tout. A partir de 10 mesures, on obtient par exemple :

```
sd(runif(10))
```

```
## [1] 0.2363325
```

Le théorème de la limite centrale nous dit que la précision de la mesure moyenne augmente avec la racine carrée du nombre de mesures : en quadruplant le nombre de mesures, l'intervalle de confiance est divisé par deux (la précision est deux fois meilleure). On peut vérifier ce résultat par simulation, en répétant un grand nombre de fois (10000) notre mesure moyenne (obtenue toujours en moyennant 1000 mesures) pour observer sa distribution :

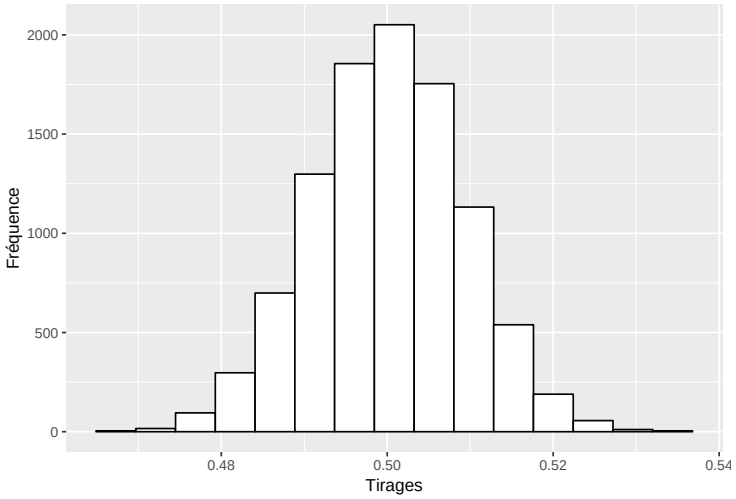

FIGURE 3: Histogramme des fréquences de la mesure moyenne (moyenne de 1000 tirages d'une loi uniforme entre 0 et 1) répétée 10000 fois. On observe une distribution en cloche typique d'une loi normale. La dispersion de la mesure moyenne est faible : aucune simulation n'a donné de résultat inférieur à 0,47 ni supérieur à 0,53, à comparer avec les histogrammes de la figure
reffig :hist10.

```
NbTirages <- 1000
NbRepetitions <- 10000
TiragesM <- replicate(NbRepetitions, mean(runif(NbTirages)))
ggplot(data.frame(TiragesM), aes(TiragesM)) + geom_histogram(bins = nclass.Sturges(TiragesM),
    color = "black", fill = "white", boundary = 0) +
    labs(x = "Tirages", y = "Fréquence")
```

Les résultats peuvent être affichés sous forme de densité de probabilité pour être comparés à une distribution normale (Figure 4).

```
fd <- function(NbTirages, TiragesM) {
    thePlot <- ggplot() + # Affichage graphique de la densité de probabilité
    geom_density(aes(TiragesM), data.frame(TiragesM)) +
        # Tracé de l'intervalle de confiance
    geom_vline(xintercept = quantile(TiragesM, probs = 0.025),
        lty = 2) + geom_vline(xintercept = quantile(TiragesM,
        probs = 0.975), lty = 2)
    # Calcul de la densité d'une loi normale entre 0 et
    # 1 par pas de 0.001
    StdErr <- 1/sqrt(12 * NbTirages)
    Normale <- dnorm(seq(0, 1, 0.001), mean = 0.5,
        sd = StdErr)
    thePlot <- thePlot + # Tracé de la densité de la loi normale
    geom_line(aes(x = x, y = y), data.frame(x = seq(0,
        1, 0.001), y = Normale), col = "red") + # Tracé des quantiles de la loi normale
    geom_vline(xintercept = qnorm(p = 0.025, mean = 0.5,
        sd = StdErr), lty = 2, col = "red") + geom_vline(xintercept = qnorm(p = 0.975,
        mean = 0.5, sd = StdErr), lty = 2, col = "red") +
        labs(x = "Tirage", y = "Densité") + xlim(0.5 -
```

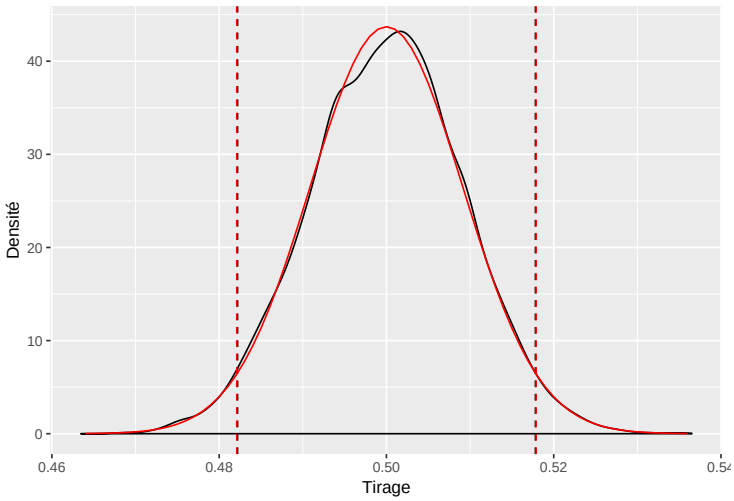

FIGURE 4: Distribution de la mesure moyenne représentée sous la forme d'une densité de probabilité (lissage de l'histogramme de la figure précédente). La densité de probabilité de la loi normale vers laquelle la distribution converge est représentée en rouge. Les quantiles sont tracés en pointillés noirs pour la mesure moyenne (résultat des simulations) et rouge pour la loi normale (valeur théorique).

```
        4 * StdErr, 0.5 + 4 * StdErr)
    return(thePlot)
}
fd(NbTirages, TiragesM)
```

L'intervalle de confiance empirique (pointillés noirs, obtenu à partir des simulations) est très proche de l'intervalle de confiance théorique (pointillés rouges, calculé pour la loi normale : $[0,482; 0,518]$).

En augmentant le nombre de mesures, par exemple à 100000, c'est-à-dire 100 fois plus que précédemment, la distribution s'approche encore de la loi normale et l'intervalle de confiance est encore divisé par 10.

En diminuant la précision de chaque mesure, le résultat ne change pas : répétons la simulation en limitant la précision des mesures à un seul chiffre significatif, c'est-à-dire en arrondissant les simulations à un chiffre après la virgule (les valeurs simulées ne peuvent être que 0, 0,1, 0,2, ..., 0,9, 1). La ligne de code qui réalise les tirages devient (Figure 5) :

```
fd(NbTirages, replicate(NbRepetitions, mean(round(runif(NbTirages),
    1))))
```

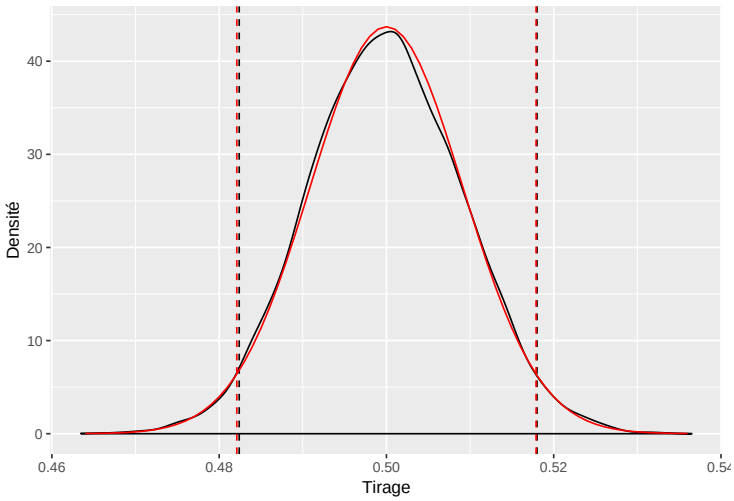

FIGURE 5: Simulation identique à celle de la figure précédente en limitant la précision de la mesure (chaque valeur simulée est arrondie à un seul chiffre significatif). Les mesures sont répétées 1000 fois, les erreurs d'arrondi se compensent et l'estimation de la moyenne est aussi bonne qu'avec les valeurs non arrondies.

4 Dimensionnement

Dans la réalité, on cherchera souvent à dimensionner une expérience. A partir de 10 mesures, on estime rapidement la moyenne et l'écart-type :

```
Tirages <- runif(10)
mean(Tirages)
```

```
## [1] 0.6602652
```

```
sd(Tirages)
```

```
## [1] 0.2165984
```

On trouve ici par exemple 0,53 pour la moyenne et 0,27 pour l'écart-type. Dans la réalité, on ne connaît pas la loi des mesures mais cette estimation rapide permet de décider de l'effort de mesure en fonction de la précision choisie.

Supposons qu'on veuille estimer la valeur avec une précision de $\pm 10\%$, c'est-à-dire un intervalle de confiance de l'ordre de 0,05. On note n le nombre de mesures nécessaires. On sait que l'intervalle de confiance (à 95%) d'une loi normale est de ± 2 écarts-types divisés par $\sqrt{n}$. On calcule donc facilement le nombre de mesures nécessaires :

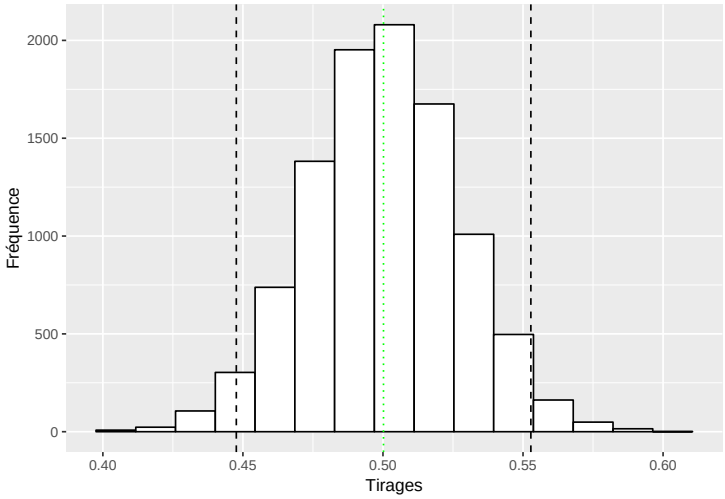

FIGURE 6: Distribution d'une mesure moyenne réalisée sur 116 individus. La valeur moyenne est tracée en vert (0,5), l'intervalle de confiance empirique à 95% est en pointillés noirs.

$$0,05 > 2 \times 0,27/\sqrt{n} \Leftrightarrow n > (2 \times 0,27/0,05)^2 \approx 116$$

En faisant 116 mesures, on s'attend d'après notre première estimation à trouver un résultat de l'ordre de $0,53 \pm 0,05$ au seuil de confiance de 95% (il y a 5% de chances que notre mesure moyenne tombe par hasard en dehors de cet intervalle). Vérifions par simulations, en répétant le script des figures précédentes avec 116 tirages (Figure 6) :

```
NbTirages <- 116
NbRepetitions <- 10000
TiragesM <- replicate(NbRepetitions, mean(runif(NbTirages)))
ggplot(data.frame(TiragesM), aes(TiragesM)) + geom_histogram(bins = nclass.Sturges(TiragesM),
    color = "black", fill = "white", boundary = 0) +
    # Tracé de l'intervalle de confiance et de la
# moyenne
geom_vline(xintercept = quantile(TiragesM, probs = 0.025),
    lty = 2) + geom_vline(xintercept = quantile(TiragesM,
    probs = 0.975), lty = 2) + geom_vline(xintercept = mean(TiragesM),
    lty = 3, col = "green") + labs(x = "Tirages", y = "Fréquence")
```

La répétition des simulations montre que la moyenne est égale à 0,5 et pas 0,53, mais l'intervalle de confiance recherché est bien atteint : environ 95% des mesures moyennes sont entre 0,45 et 0,55.

La précision peut être améliorée de deux façons, à ne pas confondre : - On peut vouloir que l'intervalle de confiance soit 10 fois plus petit : $\pm 0,005$ dans notre exemple. Alors, il faudra 100 fois plus de mesures. - On peut décider aussi d'augmenter le seuil de confiance : un intervalle de confiance à 95% signifie que

la mesure moyenne sera plus grande ou plus petite dans 5% des cas. Ce seuil est arbitraire. Il peut être trop faible (pour un médicament par exemple, on peut avoir envie d'être confiant à 99% ou plus dans le résultat). Au seuil de 99%, l'intervalle de confiance n'est plus de ± 2 écarts-types divisés par $\sqrt{n}$, mais de presque ± 3 écarts-types divisés par $\sqrt{n}$. La valeur exacte est celle du quantile de la loi de Student[4] et dépend du seuil de confiance et de n (elle vaut précisément 1.96 et non 2 dans les simulations ci-dessus). Sa valeur est calculable dans R :

```
# 95 pour cent
qt(1 - (1 - 0.95)/2, 1000 - 1)
```

```
## [1] 1.962341
```

```
# 99 pour cent
qt(1 - (1 - 0.99)/2, 1000 - 1)
```

```
## [1] 2.58076
```

La syntaxe étrange de la fonction *qt()* est due à ses usages multiples : les anglo-saxons raisonnent plutôt en terme de risque (5%) que de confiance (95%), et ce risque est ici divisé par 2 parce qu'on élimine que 2,5% des observations les plus grandes et les plus petites. L'autre paramètre est le nombre de degrés de liberté[5], égal au nombre d'observations, 1000, moins 1 (la moyenne est fixée).

5 Conclusion

Répéter des mesures permet d'améliorer la précision de l'estimation de leur moyenne : la loi des grands nombres dit que toutes les caractéristiques de la distribution empirique se rapprochent de celle de la loi théorique (on l'a vu pour l'histogramme, la moyenne et l'écart-type).

Quand le nombre de répétitions est assez grand, le théorème de la limite centrale dit que la moyenne des mesures suit une loi qui converge vers une loi normale dont l'écart-type diminue proportionnellement à la racine carrée du nombre d'observations. Il est donc possible de calculer un intervalle de confiance pour la mesure moyenne : au seuil de 95%, il est approximativement de ± 2 écarts-types divisés par la racine carrée su nombre d'observations.

Pour améliorer la précision de l'estimation, il est nécessaire de multiplier les mesures. La précision de chaque mesure n'a que peu d'importance, tant que les erreurs se compensent. On préférera donc faire beaucoup de mesures avec un appareil peu précis mais rapide que peu de mesures très précises.

Dans la réalité, l'écart-type de la mesure est inconnu a priori. Il peut être estimé sur un échantillon de petite taille, et le nombre de mesures nécessaires pour atteindre une précision choisie peut être calculé.

Références

[4] http ://fr.wikipedia.org/wiki/Loi_de_Student
[5] http ://fr.wikipedia.org/wiki/Degr%C3%A9_de_libert%C3%A9_(statistiques)