Statistiques avec R

SCI 1018

Statistiques descriptives

Marc J. Mazerolle

Département des sciences du bois et de la forêt, Université Laval

Avec révisions mineures de

Élise Filotas, *Département science et technologie, Université TÉLUQ*, et

Marc-Olivier Martin-Guay, *Département des sciences biologiques, Université du Québec à Montréal*

Table des matières

Notez bien. Le contenu théorique de chaque leçon est présenté dans un document comme celui-ci et intègre souvent du code en langage de programmation `R`. Nous sommes conscients que la convention en français est d'utiliser la virgule pour indiquer la décimale. Toutefois, nous utilisons systématiquement le point pour désigner la décimale dans le texte. Ce choix vient du fait que la syntaxe de `R` utilise le point comme décimale – à la fois pour la saisie de valeurs numériques et pour l'affichage des sorties d'analyses. Ainsi, l'usage du point uniformisera le texte et nous sommes d'avis que cette décision facilitera sa compréhension.

1 À quoi servent les statistiques ?

Les statistiques constituent une partie importante de la démarche scientifique. Elles s'appliquent à des domaines aussi variés que l'écologie, le génie, la psychologie, la médecine, et les sciences sociales. Les objectifs de l'analyse statistique sont les suivants :

- Estimer des paramètres. Par exemple, la question « Quel est le taux d'obésité au Québec ? » est un problème d'estimation. C'est-à-dire que nous cherchons à trouver la vraie valeur de cette densité.
- Tester des hypothèses. Par exemple, en posant la question « Les ours des aires protégées du Québec ont-ils une plus grande masse que leurs homologues à l'extérieur des aires protégées ? » Ici, nous voulons déterminer si la différence observée dans les deux groupes est due au hasard ou à un réel effet.
- Inférer les résultats dans un contexte plus général (où inférer signifie « tirer des conclusions »). C'est ce que font, par exemple, régulièrement les maisons de sondage d'opinions politiques auprès de la population.
- Faire des prédictions. La prédiction consiste à construire un modèle et l'utiliser pour prédire le comportement d'une variable d'intérêt. On peut utiliser la prédiction en médecine, météorologie ou en toxicologie, par exemple.

Définissons maintenant certains termes et concepts en statistique.

2 Paramètre vs statistique

Le **paramètre** est un concept important. Il désigne une valeur numérique inconnue qui caractérise une population d'intérêt. Par exemple, la taille moyenne en cm des résidents de l'île de Montréal est une valeur inconnue (mais qui existe). C'est-à-dire qu'il serait possible de calculer cette valeur si on mesurait chaque individu de cette région. Un paramètre est habituellement représenté par une lettre grecque (μ, σ). Si la taille moyenne des résidents était de 1.91 m, on écrirait $\mu = 1.91$ m.

À l'opposé, une **statistique** est une quantité qui peut être calculée à partir des données d'un échantillon. Par exemple, si nous désirons calculer la taille moyenne des résidents de l'île de Montréal, nous pourrions le faire en mesurant la taille de 100 résidents de l'île. Une statistique est normalement désignée par une lettre romaine (s, sd, $\bar{x}$).

La **population statistique** est l'ensemble des éléments sur lesquels on veut baser nos conclusions (taille de tous les résidents de Montréal). On ne connaît pas la taille moyenne des gens de cette population. Il existe deux options afin d'obtenir de l'information sur cette moyenne :

- mesurer la taille de chaque résident de Montréal (peu pratique et logistiquement difficile) ;
- utiliser un **échantillon** construit à partir de tailles d'individus sélectionnés aléatoirement dans la population de Montréal.

Pour faire une analogie, l'échantillon est à la population, ce que la statistique est au paramètre. Si nous poursuivons avec notre exemple d'échantillon de 100 résidents de Montréal (100 observations), la valeur numérique obtenue constituera une estimation de la taille moyenne (μ) des résidents de Montréal. L'estimation est une valeur possible que peut prendre un paramètre. Pour récapituler, on infère sur la population à partir d'un échantillon. Si la moyenne de l'échantillon est de 1.7 m ($\bar{x} = 1.7$ m), on peut dire que 1.7 est une estimation de la moyenne de la population μ. Bref, on peut tirer des conclusions sur la population à partir d'un échantillon qui provient de cette même population.

Afin de faire une bonne estimation, l'échantillon doit être aléatoire et représentatif de la population. Dans un échantillon aléatoire, chaque élément de la population a une chance égale d'être inclus dans l'échantillon. Si on sélectionne aléatoirement 100 résidents de Montréal pour estimer la taille moyenne des individus dans la population et que, par malchance, tous les résidents sélectionnés proviennent du même quartier, l'échantillon ne sera pas représentatif de la population.

3 Mesures de la tendance centrale

Certaines mesures décrivent la valeur autour de laquelle se concentrent la plupart des observations d'un échantillon ou d'une population. On parle alors de **tendance centrale** ou de **paramètres de position**. On peut estimer ces paramètres à partir d'un échantillon. La **moyenne arithmétique** est un exemple de ce genre de mesure :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

où x_i correspond à la valeur i de la variable x et n correspond au nombre d'observations. À noter que Σ (la lettre grecque *sigma*) indique la somme de toutes les observations de $i = 1$ jusqu'à n. De façon plus générale, on appelle **estimateur** une formule ou équation utilisée pour estimer une certaine valeur, alors que l'estimation est le résultat de l'estimateur.

Exemple 1.1 Lors d'une expérience sur la taille de poulets d'élevage deux semaines après leur naissance, on obtient les valeurs suivantes en cm : 12.3, 4.2, 5.9, 9.1, 3.3, 5.1, 7.3, 3.8, 8.0, 6.1. Le calcul de la moyenne arithmétique se fait comme suit :

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} = \frac{12.3 + 4.2 + 5.9 + \ldots + 8.0 + 6.1}{10} \\ \bar{x} &= 6.51 \end{aligned}$$

Ainsi, la moyenne arithmétique de cet échantillon est de 6.51 cm.

L'estimateur de $\bar{x}$ est un estimateur non-biaisé de μ si :

- les observations sont effectuées sur des individus sélectionnés aléatoirement ;

- les observations sont indépendantes ;
- les observations de la variable décrivant la population suivent une distribution normale (cette distribution sera montrée dans la section 7 de cette leçon).

La **moyenne géométrique** est une autre mesure de tendance centrale, particulièrement appropriée pour décrire des processus multiplicatifs[1]. Un processus multiplicatif donne lieu à des effets qui peuvent croître de façon exponentielle et, par le fait même, créer des valeurs extrêmes parmi les observations :

$$\begin{aligned} \bar{x}_{géom} &= \sqrt[n]{\prod_{i=1}^{n} x_i} \\ \bar{x}_{géom} &= e^{\frac{\sum_{i=1}^{n} \ln(x_i)}{n}} \end{aligned}$$

Exemple 1.2 Disons qu'après un décompte d'insectes ravageurs sur 5 quadrats[2] dans un champ agricole où ces insectes peuvent proliférer de façon exponentielle, on observe les abondances suivantes : 10, 1, 1000, 1, 10.

$$\begin{aligned} \bar{x}_{géom} &= \sqrt[5]{10 \cdot 1 \cdot 1000 \cdot 1 \cdot 10} \\ \bar{x}_{géom} &= 10 \end{aligned}$$

La moyenne géométrique de ces valeurs nous donne 10, alors que la moyenne arithmétique nous donne 204.4. La valeur 1000 se démarque nettement des autres et exerce une influence démesurée sur la moyenne arithmétique, et dans ce cas, la moyenne géométrique est un meilleur estimateur de la tendance centrale.

1. Dans un processus multiplicatif, une variable a un effet multiplicatif sur une variable réponse. Par exemple, si on remarque que la croissance de semis à concentration modérée d'engrais est 2.5 fois plus élevée qu'à concentration faible, la concentration a un effet multiplicatif sur la croissance.

2. Un quadrat est une unité spatiale de dimension donnée (1 m × 1 m, 10 m × 10 m), disposée dans un site d'étude sur laquelle on fait des mesures en écologie, ici un décompte d'insectes.

La **moyenne harmonique** peut s'appliquer à des taux (p. ex., vitesses) :

$$\bar{x}_{harm} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Exemple 1.3 Par exemple, disons que nous avons suivi un ours noir par télémétrie. L'ours a parcouru un segment de 2 km à une vitesse de 1 km/h, un deuxième segment de 2 km à une vitesse de 2 km/h, un troisième segment de 2 km à une vitesse de 4 km/h et un dernier segment de 2 km à une vitesse de 1 km/h. Quelle est la vitesse moyenne de l'ours ?

On pourrait utiliser la moyenne harmonique pour résoudre le problème. Sachant que *vitesse* = *distance/temps*, nous pouvons déterminer la distance totale parcourue : $4 \cdot 2\,\text{km} = 8\,\text{km}$. On peut ensuite évaluer le temps mis à parcourir ces 8 km :

- 1er segment : 2 km * 1h/1 km = 2 h
- 2e segment : 2 km * 1h/2 km = 1 h
- 3e segment : 2 km * 1h/4 km = 0.5 h
- 4e segment : 2 km * 1h/1 km = 2 h

Le temps total est 5.5 h. Nous pouvons calculer la vitesse moyenne :

$$\begin{aligned} \text{vitesse moyenne} &= 8\,\text{km}/5.5\,\text{h} \\ \text{vitesse moyenne} &= 1.45\,\text{km/h} \end{aligned}$$

C'est exactement ce que nous donne la moyenne harmonique :

$$\begin{aligned} \bar{x}_{harm} &= \frac{4}{1/1 + 1/2 + 1/4 + 1/1} \\ \bar{x}_{harm} &= 1.45 \end{aligned}$$

À noter que la moyenne harmonique s'applique à des vitesses si elles sont mesurées sur une même distance. Si les distances diffèrent, nous devrons utiliser une version pondérée de la moyenne harmonique.

Les trois types de moyennes sont reliées par la relation suivante :

$$\bar{x}_{harmonique} < \bar{x}_{géom} < \bar{x}$$

Si les observations sont égales ($x_1 = x_2 = x_3 \ldots = x_n$), nous obtenons :

$$\bar{x}_{harmonique} = \bar{x}_{géom} = \bar{x}$$

Il existe d'autres mesures de tendance centrale, notamment la **médiane** qui se définit comme étant la valeur qui sépare les observations en deux groupes égaux (50 % des valeurs $<$ médiane, 50 % des valeurs $>$ médiane). Pour obtenir la médiane, il suffit d'ordonner les observations dans une liste croissante. La médiane est donc l'observation au milieu de cette liste si le nombre d'observations est impair, ou la moyenne entre les deux observations au milieu de la liste si ce nombre est pair. En présence de données normales, la médiane et la moyenne sont proches. La médiane est peu influencée par la présence de valeurs extrêmes (valeurs très grandes ou très faibles), alors que la moyenne est très sensible à la présence de valeurs extrêmes.

Exemple 1.4 Dans une expérience sur le temps de survie d'insectes exposés à un insecticide, on obtient les valeurs (en secondes) 1.1, 1.2, 1.3, 1.6, 3.2, 2.4, 5.2. La moyenne arithmétique de cet échantillon est de 2.29 secondes et la médiane est de 1.6 secondes. Si l'on ajoute une dernière observation dont la valeur extrême est 40 secondes, la moyenne arithmétique sera alors de 7 secondes et la médiane de 2 secondes. On constate que la médiane est beaucoup moins sensible à l'ajout de la valeur extrême, ce qui n'est pas le cas de la moyenne arithmétique.

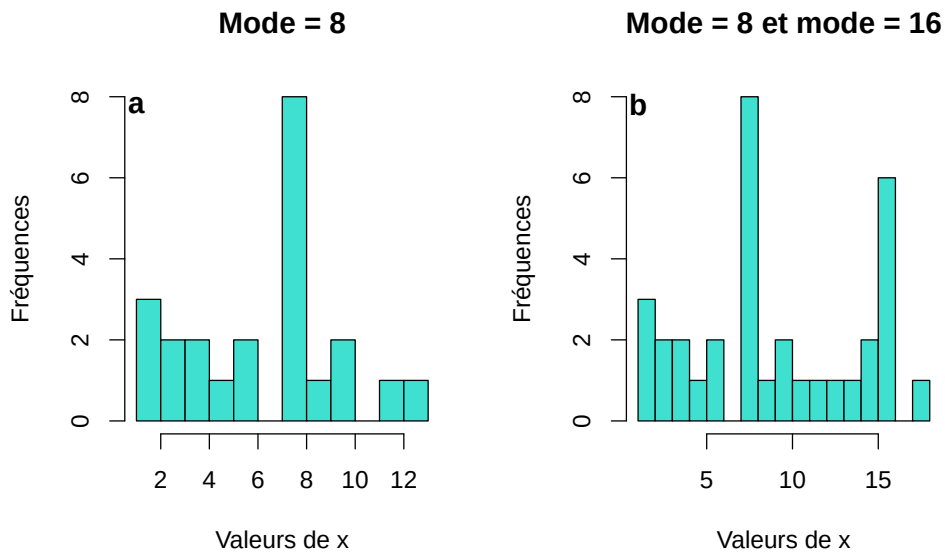

FIGURE 1 – Histogramme illustrant une distribution unimodale (a) et bimodale (b).

Le **mode** permet aussi de caractériser la tendance centrale, car il donne la ou les valeurs qui reviennent le plus souvent dans l'échantillon (fig. 1). Par exemple, si, dans un échantillon, on obtient les valeurs 12, 12, 12, 12, 12, 3, 3, 3, 3, 3, 3, 1, 2, 14, 15, 16, 21, 32, on dira qu'il y a deux modes (12 et 3).

4 Mesures de dispersion

Certaines mesures décrivent plutôt l'étendue de la variabilité des données. On parle alors de **mesures de dispersion** ou de **paramètres de variabilité**. Plus la variabilité augmente, plus l'incertitude quant à la valeur des paramètres estimés à partir de données d'un échantillon augmente. Un niveau d'incertitude plus élevé augmente la difficulté de trouver des différences et de tester des hypothèses. L'**étendue** (*range*) est la mesure de dispersion la plus simple. Il s'agit de la différence entre la valeur minimale et la valeur maximale des observations.

4.1 Somme des carrés des erreurs

La **somme des carrés des erreurs** (*sum of squared errors, SSE*) donne le carré de la différence entre chaque observation et la moyenne de l'échantillon :

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Cette mesure de variabilité est l'une des plus communes, et peut prendre des valeurs ≥ 0 (le carré assure des valeurs positives). Plus cette valeur est grande, plus il y a de variabilité dans les données (i.e., les observations sont plus éloignées de la moyenne).

Exemple 1.5 Un échantillon de 6 longueurs de tige d'une plante ligneuse donne 1.3 m, 4.5 m, 4.1 m, 2.1 m, 5.0 m, et 1.9 m, il s'ensuivra que $\bar{x} = 3.15$ m et que $SSE = (1.3 - 3.15)^2 + (4.5 - 3.15)^2 + \ldots + (1.9 - 3.15)^2 = 12.24\ \text{m}^2$. Une propriété importante de la SSE est qu'à chaque nouvelle observation ajoutée, elle augmente (pourvu que $x_{nouvelle} \neq \bar{x}$). Si on ajoute une septième valeur de 2.6 à notre échantillon de longueurs de tige présenté ci-haut, la moyenne arithmétique devient 3.07 et la SSE s'élèvera à 12.49 m^2.

Une meilleure mesure de dispersion devrait tenir compte de la taille de l'échantillon. Mais avant d'aller plus loin, allons visiter le concept de **degrés de liberté** (*degrees of freedom, df*), un concept souvent nébuleux que nous tenterons d'éclaircir ici. On peut voir les degrés de liberté comme étant la taille de l'échantillon corrigée pour le nombre de paramètres estimés. On peut obtenir cette valeur en soustrayant le nombre de paramètres estimés p de la taille d'échantillon n (i.e., $n - p$). Clarifions avec un exemple.

Exemple 1.6 Imaginez qu'on ait un échantillon de 5 observations dont on ne connait rien. Ces 5 observations pourraient prendre n'importe quelle valeur. Le degré de liberté est donc 5 ($df = 5$). Imaginez maintenant qu'on connaisse un paramètre de cet échantillon (p. ex., $\bar{x} = 7$). On réduit la liberté des valeurs que peuvent prendre ces 5 observations. En effet, disons que les valeurs aient été déterminées pour 4 des observations et que la moyenne est connue, la dernière observation est obligée de prendre une valeur en particulier. Avec un paramètre connu, le degré de liberté est donc 4 ($df = 5 - 1 = 4$).

4.2 Carré moyen (variance)

Comme nous l'avons mentionné plus tôt, la somme des carrés des erreurs (SSE) augmente avec la taille de l'échantillon. Une meilleure mesure devrait tenir compte de la taille d'échantillon. Le **carré moyen** (*mean square, mean squared error, MSE*) est une telle mesure de dispersion :

$$MSE = \frac{SSE}{df} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

À noter que le dénominateur correspond aux degrés de liberté, ici $n-1$, puisque nous avons estimé la moyenne arithmétique μ à l'aide de $\bar{x}$ pour trouver la SSE. Ce carré moyen est en fait la **variance** de l'échantillon, $s^2 = MSE$. Cette relation est importante et nous reviendrons sur cette notion lors de la leçon sur l'analyse de variance. On peut donc estimer la variance de la population à partir d'un échantillon en utilisant l'équation :

$$\begin{aligned} s^2 &= MSE = \frac{SSE}{df} \\ s^2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \end{aligned}$$

L'**écart-type** (s) est simplement la racine carrée de la variance, et il indique la variabilité dans les données. La variance dépend énormément de la taille de l'échantillon. L'estimation devient difficile lorsqu'on a peu d'observations dans l'échantillon. Illustrons avec un exemple.

Exemple 1.7 Utilisons une petite simulation à l'aide de `R` pour générer des données provenant d'une population avec des caractéristiques connues, soit une population normale avec une moyenne de 10.1 ($\mu = 10.1$) et une variance de 4 ($\sigma^2 = 4$). Nous allons sélectionner aléatoirement trois observations provenant de cette population afin de constituer un échantillon de $n = 3$.

À partir de cet échantillon, nous pouvons calculer une variance qui sera une estimation de la vraie valeur. Nous estimerons σ^2 à l'aide de l'estimateur de la variance (s^2) d'un échantillon. Afin d'obtenir une meilleure idée de la performance de l'estimation de la variance, nous allons ensuite répéter l'exercice pour 29 autres échantillons de $n = 3$ tirés de la même population, et calculer la variance de chaque échantillon de taille 3. Par la suite, nous ferons de même pour 30 échantillons

constitués de 4 observations, 30 échantillons de 5 observations, . . . , 30 échantillons de 49 observations et 30 échantillons de 50 observations (fig. 2).

On remarque que l'estimation de la variance est parfois très loin de la vraie valeur de 4, particulièrement pour les très petits échantillons ($n \leq 10$). On obtient de meilleures estimations pour de plus grands échantillons, particulièrement au-delà de 30. C'est une des raisons pour laquelle on considère généralement un échantillon de 30 observations comme ayant une taille suffisante – il permet de bien estimer la variance. On comprend rapidement que l'utilisation d'un petit échantillon peut nous amener loin de la vraie valeur de la variance. Mais pourquoi s'intéresser autant à la variance ?

La variance est une quantité importante en statistiques, puisqu'elle est requise pour construire des mesures de précision (p. ex., intervalles de confiance) et pour tester des hypothèses (p. ex., test t). Un petit échantillon peut produire une estimation très loin de la vraie valeur de la variance et invalider les conclusions d'une analyse statistique. Tel qu'illustré dans l'exemple 1.7, l'estimation de la variance s'améliore avec la taille de l'échantillon. Ce qui nous mène à visiter les concepts de **précision** et **d'exactitude**.

4.3 Précision vs exactitude

La réalisation d'une expérience, impliquant l'échantillonnage des observations et l'estimation des quantités, s'apparente à un archer qui lance une flèche sur une cible, où la flèche correspond à une expérience et le point sur la cible correspond à une estimation. On veut que la flèche se rende le plus près du centre de la cible (c.-à-d., une bonne estimation), mais on veut que les flèches ne soient pas trop éloignées les unes des autres (c.-à-d., une bonne précision). En d'autres termes, un archer est précis si toutes ses flèches tombent très près du même point sur la cible (fig. 3a, c), ou encore il peut manquer d'exactitude lorsque ses flèches

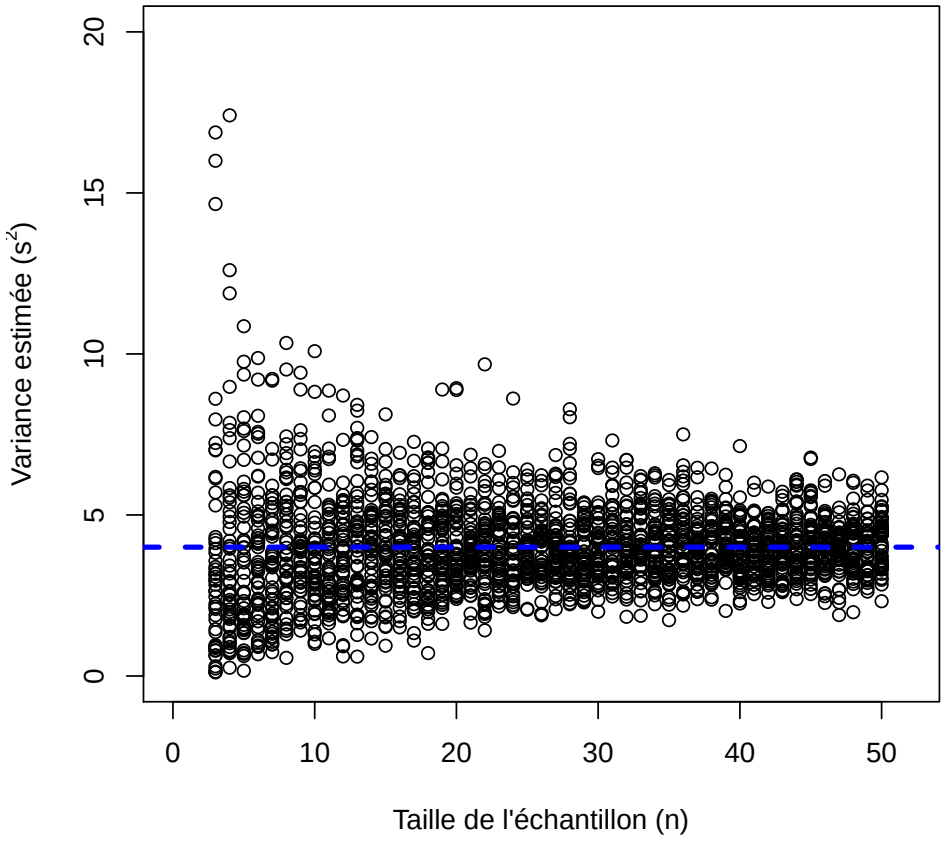

FIGURE 2 – Effet du nombre d'observations sur l'estimation de la variance. À noter que la ligne pointillée représente la vraie valeur de la population ($\sigma^2 = 4$) à partir de laquelle les observations ont été sélectionnées aléatoirement.

sont loin du centre de la cible (fig. 3c, d). Le meilleur des scénarios est un tir précis et exact (fig. 3a), et le pire est un tir ni précis, ni exact (fig. 3d). Le tir exact mais peu précis implique que l'estimation varie beaucoup d'un échantillon à l'autre (fig. 3b), et cette variation n'est pas souhaitable.

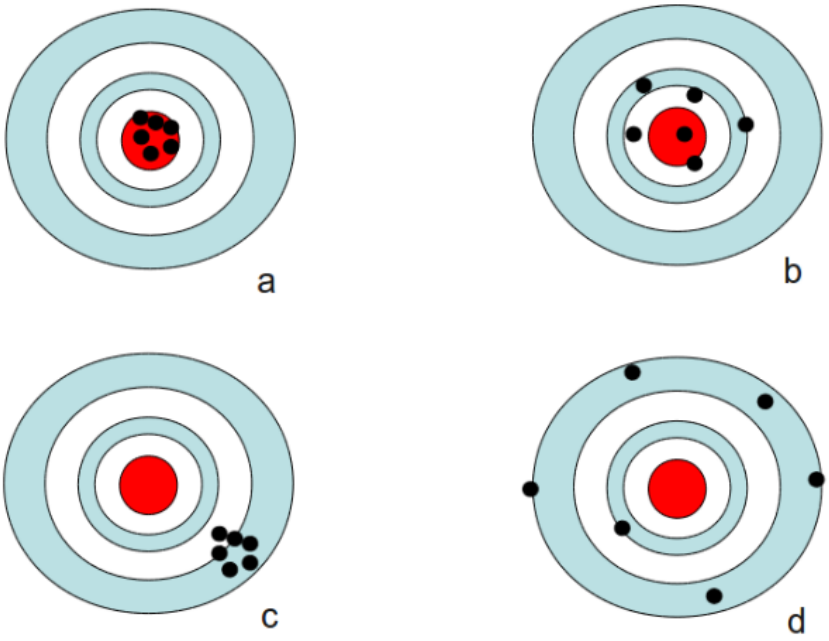

FIGURE 3 – Utilisation de cibles pour expliquer le concept de précision et d'exactitude avec quatre archers dans une compétition. Si tous les points se trouvent au centre et très près les uns des autres, l'archer est précis et exact (a), alors que si les points sont dans la région centrale mais éloignés les uns des autres, l'archer est exact mais peu précis (b). À l'opposé, si les points sont très près les uns des autres et loin du centre, l'archer est précis mais manque d'exactitude (c), tandis que dans le dernier scénario l'archer n'est ni précis ni exact (d).

L'exactitude peut être vue comme un terme qualitatif. La valeur qui quantifie la déviation entre les estimations et la valeur réelle du paramètre s'appelle biais. Plus formellement, on appelle **biais** la différence entre la **valeur attendue** d'une estimation et la valeur réelle qu'on désire estimer :

$$biais = E(\hat{\theta}) - \theta$$

où θ est la valeur réelle du paramètre de la population, $\hat{\theta}$ correspond à l'estimation d'un paramètre obtenu avec un seul jeu de données, et $E(\hat{\theta})$ est la valeur attendue des estimations du paramètre. La valeur attendue est en fait la moyenne d'une série d'estimations $\hat{\theta}$ obtenues à partir de plusieurs échantillons de taille égale provenant de la même population. Pour poursuivre notre analogie des archers, la valeur attendue correspond à la moyenne des positions des flèches sur la cible. Le biais exprime la tendance des différences entre les valeurs estimées d'un paramètre et la vraie valeur de ce paramètre. Lorsque le biais est de 0, on dit que l'estimateur est non biaisé (p. ex., l'estimateur de la moyenne arithmétique, $\bar{x}$, sous certaines conditions).

Si on développait une mesure d'imprécision ($1/\textit{précision}$), on s'attendrait à ce qu'elle augmente proportionellement ($\propto$[3]) avec la variance :

$$\textit{imprécision} \propto s^2$$

Par contre, on s'attendrait à ce que l'imprécision diminue avec la taille de l'échantillon :

$$\textit{imprécision} \propto \frac{s^2}{n}$$

Une mesure idéale s'exprimerait dans les mêmes unités que les observations :

$$\textit{imprécision} \propto \sqrt{\frac{s^2}{n}}$$

Une telle mesure existe déjà, c'est l'erreur-type de la moyenne (SE) :

$$SE = s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$$

L'erreur-type de la moyenne d'un échantillon ($s_{\bar{x}}$ ou SE) représente l'écart-type de la

3. Le symbole $\propto$ indique la proportionnalité entre deux variables. En d'autres mots, on peut passer des valeurs d'une variable en multipliant ou divisant par une constante non nulle pour obtenir les valeurs de l'autre.

distribution des moyennes calculées à partir d'échantillons de taille identique à celle de notre échantillon. Ainsi, l'erreur-type de la moyenne nous donne une indication sur la variabilité de l'estimation de ce paramètre si on répétait l'échantillonnage. Pour clarifier, l'écart-type nous informe sur la variabilité d'un échantillon alors que l'erreur-type de la moyenne nous indique la précision avec laquelle nous avons estimé ce paramètre. Nous revisiterons ce concept lors des deux dernières leçons consacrées aux modèles de régression.

L'erreur-type nous permet de calculer des **intervalles de confiance** autour de la moyenne ou d'autres paramètres. L'intervalle de confiance est justement une autre mesure de précision autour d'une estimation. L'intervalle de confiance (IC) se définit comme étant l'intervalle à l'intérieur duquel se trouvera la moyenne de la population μ si l'on répète l'expérience un grand nombre de fois. Pour un IC à 95 %,

$$P(\bar{x} - 1.96 \cdot SE \leq \mu \leq \bar{x} + 1.96 \cdot SE) = 0.95$$

Un IC à 95 % indique que la moyenne de la population (μ) devrait se trouver 95 % du temps (c.-à-d., la probabilité est de 0.95) à l'intérieur de l'intervalle si on répète l'échantillonnage à plusieurs reprises avec le même nombre d'observations. Pour un IC donné, la moyenne de la population μ est incluse ou non[4]. L'IC est construit à partir d'un échantillon, mais concerne la moyenne μ de la population. Nous expliquerons en détail la construction et l'interprétation d'intervalles de confiance à la prochaine leçon (voir aussi la section 7 de la présente leçon pour comprendre l'origine du facteur 1.96). Pour l'instant, il suffit de réaliser qu'on peut utiliser cette intervalle pour en indiquer la précision.

4. Les gens ont souvent une interprétation bayésienne de l'intervalle de confiance en affirmant que c'est la probabilité que la moyenne soit comprise dans l'intervalle de confiance construit à partir d'un seul échantillon. Dans le cours, nous ferons uniquement appel aux statistiques classiques aussi appelées fréquentistes et nous utiliserons la définition classique de l'intervalle de confiance.

4.4 Autres mesures de dispersion

Le **coefficient de variation** (*coefficient of variation, CV*) est parfois utilisé pour représenter la variabilité :

$$CV = \frac{s}{\bar{x}} \cdot 100\,\%$$

où s représente l'écart-type de l'échantillon et $\bar{x}$ correspond à la moyenne arithmétique de l'échantillon. On remarque que le CV est le ratio entre l'écart-type et la moyenne arithmétique. Un échantillon avec un CV de 14 % varie moins qu'un autre avec un CV de 30 %.

Les **quantiles** peuvent également nous aider à représenter à quel point les données varient. Le mot « quantile » est un terme générique qui désigne une quantité qui divise les données en compartiments après qu'elles ont été mises en ordre croissant. Les **quartiles** divisent les données en quatre compartiments, les **déciles** en 10 compartiments, et les **percentiles** en 100 compartiments. Par exemple, un 90^{e} percentile de 120 g signifie que 90 % des valeurs sont inférieures à 120 g et que 10 % des valeurs sont supérieures à 120 g.

5 Variables aléatoires

On désigne **variable aléatoire** une variable dont les valeurs observées sont considérées comme résultant d'un processus aléatoire (c.-à-d., expérience aléatoire). Autrement dit, les valeurs exactes d'une variable aléatoire dans un échantillon ne peuvent être anticipées avec certitude avant de recueillir l'échantillon que nous utiliserons pour tirer des conclusions à propos de la population (p. ex., estimer un paramètre). Le tout implique une composante aléatoire. Par exemple, si on mesure la pression artérielle, le niveau de cholestérol, et le niveau d'activité (trois variables aléatoires) chez un groupe de gens sélectionnés aléatoirement, on ne peut prédire la valeur de ces trois variables chez un individu avant de les avoir mesurées.

Les variables aléatoires peuvent être **discrètes** ou **continues**. Par discrètes, on entend des variables binaires (p. ex., présence-absence, mort-vivant), catégoriques ordonnées ou non (p. ex., petit, moyen, grand ; poisson, invertébré, mammifère) ou encore des variables apparais-

sant sous forme d'entiers (le nombre d'interruption de courant dans 5 municipalités depuis le dernier mois : 0, 1, 12, 4). Le cas échéant, l'observation peut uniquement prendre des valeurs entières – on ne peut avoir dénombré 2.4 individus dans une unité d'échantillonnage ou avoir un individu au 3/4 mort.

Les variables continues sont celles qui peuvent prendre une infinité de valeurs sur un intervalle donné. La distance, la masse, le temps, la température, la longueur sont des variables pouvant être mesurées avec différentes résolutions, selon l'instrument utilisé pour effectuer la mesure. Cette infinité de valeurs possibles fait que plus un instrument a une résolution élevée (mesure avec plus de chiffres significatifs), moins il est probable que deux observations aient la même valeur dans un échantillon. Par exemple, il est probable que deux serpents aient la même longueur au dm près, mais cela est moins probable s'ils sont mesurés au cm près, et encore moins probable au mm près.

La présentation des valeurs d'une variable peut aussi varier selon le type de variable. Nous utilisons habituellement un diagramme à bâtons pour une variable discrète, alors qu'un histogramme illustre mieux les données d'une variable continue (fig. 4).

6 Loi des grands nombres et théorème de la limite centrale

Deux principes importants agissent sur l'échantillonnage (et les échantillons) et nous permettent d'analyser les données. La **loi des grands nombres** stipule que la moyenne de l'échantillon ($\bar{x}$) tend vers la moyenne de la population (μ) au fur et à mesure que la taille de l'échantillon augmente. D'où l'importance d'une bonne taille d'échantillon.

Le **théorème de la limite centrale**, quant à lui, indique que, si on prend plusieurs échantillons indépendants d'une même population, et que l'on calcule la moyenne (ou somme) de chacun, ces moyennes (ou sommes) auront une distribution normale (voir prochaine section). Grace à ce théorème, on peut effectuer des analyses à partir d'un échantillon, même si on ne

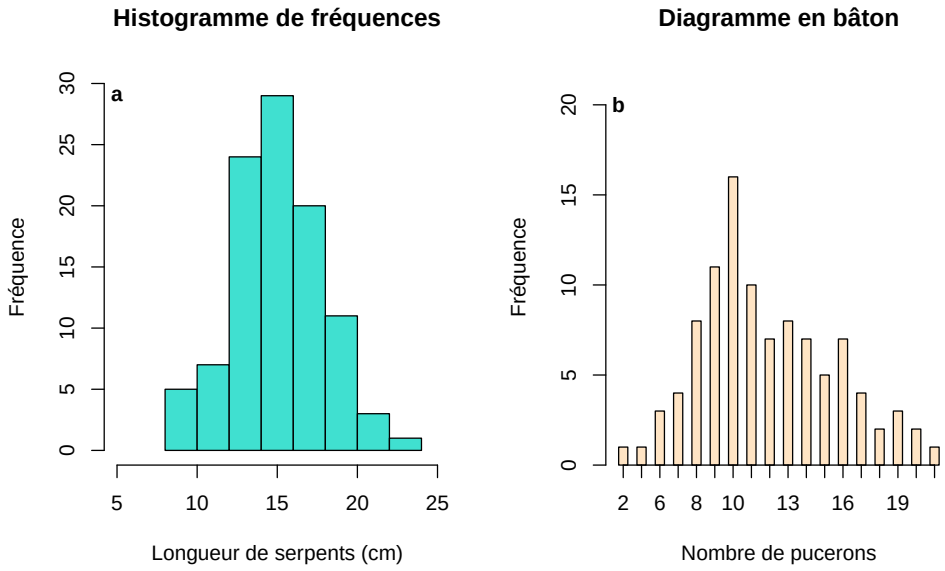

FIGURE 4 – Présentation des longueurs de 100 serpents (variable continue, a) et du nombre de pucerons dans 100 sites (variable discrète, b).

connaît pas les propriétés de la population originale. On ne peut pas généralement déterminer la normalité d'une population sans l'avoir recensée au complet. Toutefois, on peut le faire pour un échantillon qui provient de cette même population.

7 Distribution normale

La **distribution normale** (ou loi normale) est une distribution théorique centrale en statistique à la base de nombreux traitements statistiques. Découverte initialement par le mathématicien Abraham De Moivre au 17^{e} siècle et redécouverte par Karl Friedrich Gauss 100 ans plus tard, elle a été longtemps désignée sous le nom de **distribution gaussienne**.

On connaît bien les propriétés de la distribution normale et plusieurs approches utilisent cette distribution :

- les tests d'hypothèses ;
- l'estimation de paramètres par maximum de vraisemblance ;
- la construction d'intervalles de confiance.

La distribution normale se définit par la **fonction de densité de probabilité** (*probability density function, pdf*) suivante :

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} = \frac{\exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)}{\sigma\sqrt{2\pi}}$$

où x correspond à la valeur numérique d'intérêt, μ représente la moyenne de la population, σ est l'écart-type de la population, π est la constante 3.14159 ... et e est la constante 2.71828 .... Cette distribution comporte deux paramètres, μ et σ et on représente parfois cette distribution avec la notation N(μ, σ).

En mots, l'équation nous donne la densité de probabilité d'une variable qui prend la valeur x et qui provient d'une distribution normale avec une moyenne μ et un écart-type σ. Il est important de comprendre que la densité de probabilité obtenue avec cette fonction n'est pas équivalente à la probabilité d'obtenir une valeur précise. En fait, la fonction de densité de probabilité, permet de calculer la probabilité d'obtenir une valeur dans un intervalle précis en calculant l'intégrale de cette fonction pour cet intervalle (aire sous la courbe). La raison de cette distinction entre densité de probabilité et probabilité provient du fait que la probabilité d'observer une valeur spécifique (p. ex., comme la masse ou la longueur) dans une distribution continue est de 0. La valeur mesurée d'une variable continue est en réalité un intervalle qui dépend de la précision de l'instrument de mesure au mg près, au g près, ou au kg près. Par exemple, pour un serpent dont la mesure obtenue est de 102.54 cm, on obtient la probabilité suivante en utilisant l'intégrale de la fonction de densité de probabilité :

$$P(x = 102.54) = P(102.54 \leq x \leq 102.54) = \int_{102.54}^{102.54} f(x)dx = 0$$

Si la règle utilisée pour mesurer le serpent donne une précision de $\pm$ 0.01 cm, notre mesure est en fait un intervalle défini par les bornes suivantes :

$$\text{borne inférieure} = 102.54\,\text{cm} - 0.01\,\text{cm} = 102.53\,\text{cm}$$

$$\text{borne supérieure} = 102.54\,\text{cm} + 0.01\,\text{cm} = 102.55\,\text{cm}$$

La fonction de densité nous donne la densité de la distribution correspondant à la valeur x. On peut obtenir la courbe de distribution normale pour une moyenne et écart-type donnés en substituant une série de valeurs dans l'équation.

Exemple 1.8 On veut connaître la densité de probabilité associée à une masse de souris de 3.4 g dans une population de souris suivant une distribution normale ayant une moyenne de 4.1 g et un écart-type de 1.5 g (c.-à-d., N(4.1, 1.5)). Nous avons donc :

$$x = 3.4\,\text{g} \qquad f(x = 3.4|4.1, 1.5) = \frac{1}{1.5 \cdot \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{3.4-4.1}{1.5})^2} = 0.2385$$

$$\mu = 4.1\,\text{g}$$

$$\sigma = 1.5\,\text{g} \qquad .$$

On pourrait aussi déterminer la densité de probabilité associée à des souris de 8.9 g et de 10.2 g dans la même population :
Souris de 8.9 g :

$$f(x = 8.9|4.1, 1.5) = \frac{1}{1.5 \cdot \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{8.9-4.1}{1.5})^2} = 0.0016$$

Souris de 10.2 g :

$$f(x = 10.2|4.1, 1.5) = \frac{1}{1.5 \cdot \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{10.2-4.1}{1.5})^2} = 0.00007$$

On peut ensuite représenter ces valeurs sur une distribution normale avec $\mu = 4.1$ et $\sigma = 1.5$. La figure 5 illustre la densité de probabilité pour les trois souris. La courbe a été obtenue en utilisant l'équation de la densité pour une moyenne de 4.1 et un écart-type de 1.5 et en faisant varier x dans l'intervalle de 0 à 12. On remarque que la distribution est plus « dense » dans la région de 2 à 6 g. Par conséquent, on constate qu'il est plus probable de tirer au hasard une valeur dans cet intervalle (aire sous la courbe de densité plus grande) qu'une valeur à l'extérieur de cet intervalle. De plus, si on assume que chacune des mesures précédentes est associée à un intervalle qui dépend de la précision de l'instrument utilisé (p. ex., 3.4 ± 0.1 g), on peut facilement estimer que l'intervalle [3.3 g, 3.5 g] est plus probable (aire sous la courbe plus grande) que les intervalles [8.8 g, 9.0 g] et [10.1 g, 10.3 g].

La distribution normale comprend deux paramètres, μ et σ, ce qui signifie que l'on peut tracer une courbe normale dès que nous connaissons ces deux valeurs. La moyenne (μ) détermine la position (fig. 6a) et l'écart-type détermine la forme de la courbe (fig. 6b).

7.1 Caractéristiques de la distribution normale

La distribution normale est une distribution continue dans l'intervalle $[-\infty, +\infty]$. La somme de l'aire sous la courbe est 1. La distribution est symétrique autour de la moyenne μ. On sait que :

- 90 % des observations se trouvent dans l'intervalle $[\mu - 1.64\,\sigma,\ \mu + 1.64\,\sigma]$;

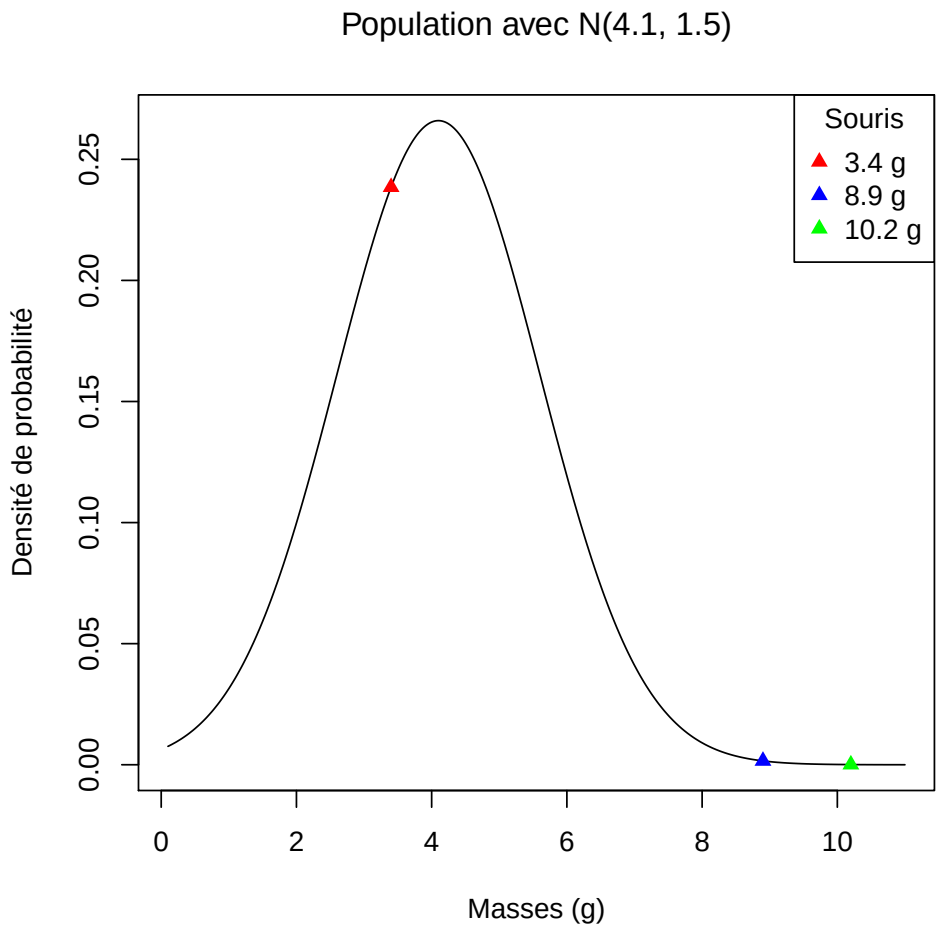

FIGURE 5 – Distribution normale d'une population de masses de souris où la moyenne est de 4.1 g et l'écart-type est de 1.5 g. À noter qu'on peut tracer la courbe de la distribution en substituant une série de valeurs de x dans la fonction de densité de probabilité pour un $\mu = 4.1$ et $\sigma = 1.5$, c.-à-d., N(4.1, 1.5).

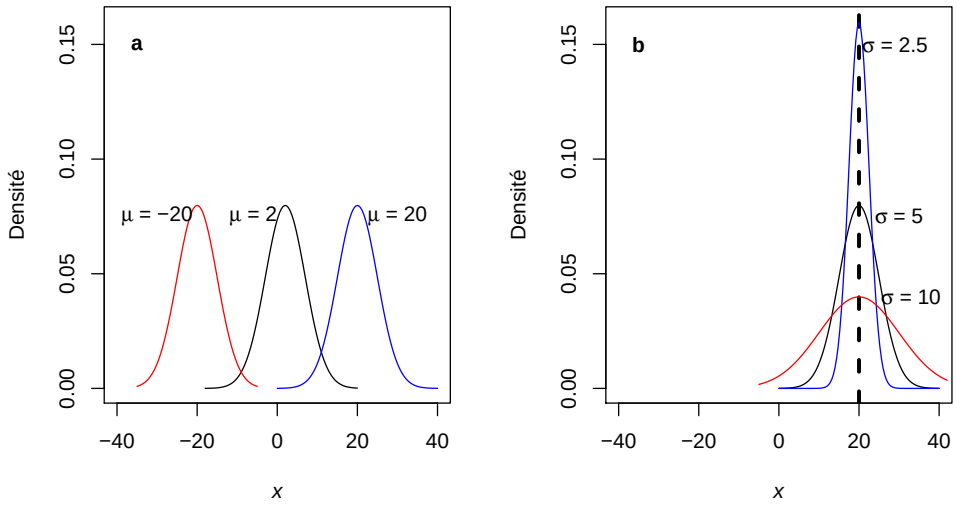

FIGURE 6 – Position de distributions normales avec même écart-type, mais différentes moyennes (a) et forme de distributions normales pour des écarts-types différents, mais une même moyenne (b).

- 95 % des observations se trouvent dans l'intervalle $[\mu - 1.96\,\sigma,\ \mu + 1.96\,\sigma]$;
- 99 % des observations se trouvent dans l'intervalle $[\mu - 2.58\,\sigma,\ \mu + 2.58\,\sigma]$.

7.2 Distribution normale centrée réduite

La distribution normale centrée réduite (*standard normal distribution*) est un cas particulier de la distribution normale où $\mu = 0$ et $\sigma = 1$ (fig. 7). **Centrer** consiste à soustraire la moyenne à chaque observation, $x_{i\,centrée} = x_i - \mu$. L'opération n'influence pas la variance, mais les observations centrées ont une moyenne de 0. Centrer et **réduire**, parfois aussi connu sous le terme **standardiser**, consiste à diviser chaque observation centrée par l'écart-type de l'échantillon, $x_{i\,centrée\,réduite} = \frac{x_i - \mu}{\sigma}$. Les observations centrées réduites ont une moyenne de 0 et un écart-type de 1. L'opération de centrer réduire est aussi appelée la transformation z ou l'écart normal, $z = \frac{x_i - \mu}{\sigma}$. Cette opération modifie l'échelle de la variable. La variable centrée réduite est exprimée en terme du nombre d'écart-types séparant chaque valeur de la

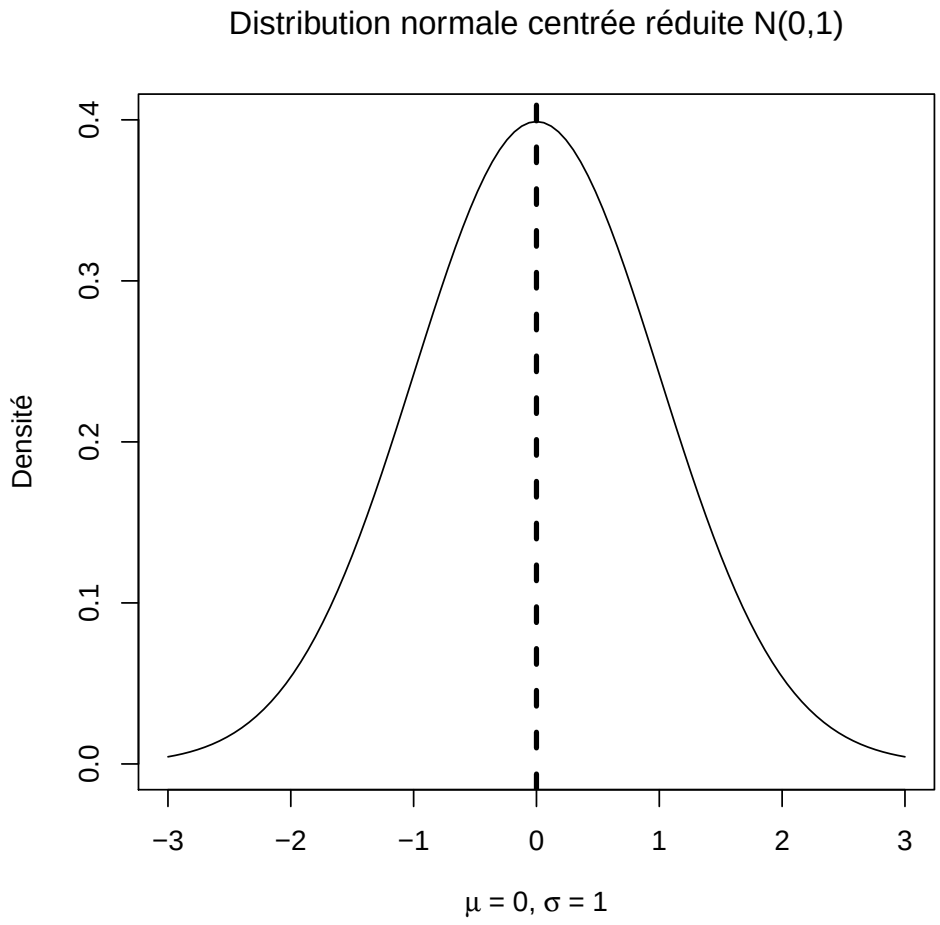

FIGURE 7 – Distribution normale centrée réduite, c.-à-d., N(0, 1).

moyenne.

Exemple 1.9 On s'intéresse à la longueur d'ailes de pucerons dans une population. Après un recensement exhaustif, on détermine que la moyenne (μ) des longueurs d'aile dans une population est de 14.2 mm et que l'écart-type (σ) est de 5.05 mm. On veut ensuite déterminer à combien d'écart-types de la moyenne se trouve une longueur d'aile de 22.6 mm chez un puceron de cette population. Nous avons donc, $z_i = \frac{x_i - \mu}{\sigma} = \frac{22.6 - 14.2}{5.05} = 1.66$. On conclut que $x_i = 22.6$ mm se

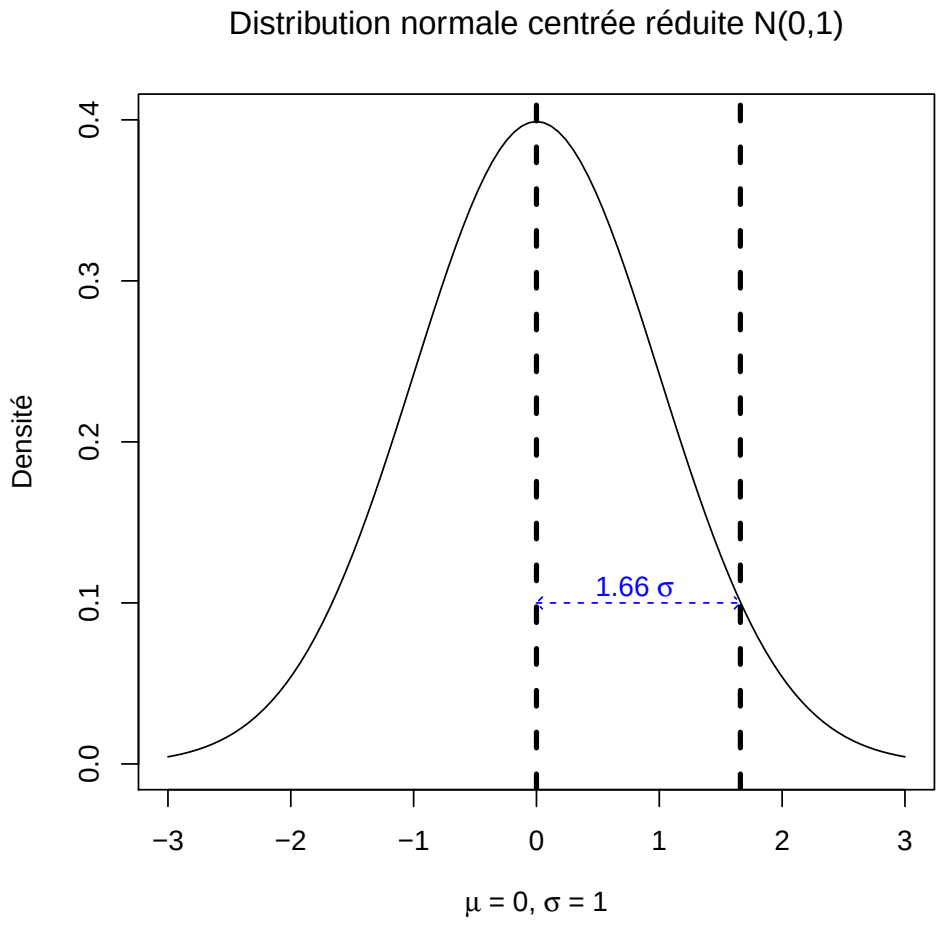

FIGURE 8 – Écart normal associé à une longueur d'ailes de puceron de 22.6 mm.

trouve à $1.66\,\sigma$ de μ (fig. 8).

7.3 Probabilités cumulatives

Les probabilités cumulatives sont beaucoup utilisées en statistique. Par exemple, on peut vouloir déterminer la probabilité d'observer un diamètre > 2.3 cm pour un arbre mesuré à une hauteur de 1 m du sol dans une forêt avec N(4, 12), c'est-à-dire, $P(\text{diamètre} > 2.3\,\text{cm})$, ou encore déterminer la probabilité d'observer une plante dont la tige mesure entre 3 et 10 cm

de longueur dans une population de plantes avec N(12.5, 3.01), $P(3\,\text{cm} < \text{longueur} < 10\,\text{cm})$.

On peut résoudre ce genre de problème à l'aide de la distribution normale ou de la distribution normale centrée réduite. La probabilité cumulative correspond à l'aire sous la courbe dans un intervalle défini par une intégrale. Par exemple, on sait que l'aire sous la courbe d'une distribution normale centrée réduite entre -1.96 et 1.96 est de 0.95 :

$$\int_{-1.96}^{1.96} f(x \mid \mu = 0, \sigma = 1)dx = 0.95$$

À noter que les probabilités cumulatives étaient autrefois obtenues à partir de tables situées en annexe de livres de statistiques. De nos jours, nous utiliserons typiquement un logiciel comme `R` pour obtenir la probabilité cumulative. Dans `R`, la fonction `pnorm( )` nous donne cette valeur et nous discuterons plus en détails de cette option dans les prochaines leçons.

Exemple 1.10 On a recensé tous les individus d'un édifice à bureaux d'une ville d'Amérique du Nord. La moyenne (μ) de la taille des individus dans cette population est de 170 cm avec un écart-type (σ) de 8 cm. Quelle est la probabilité qu'un individu soit plus petit ou égal à 160 cm dans cette population ? Pour résoudre le problème, on peut utiliser l'écart normal :

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{160 - 170}{8}$$
$$z_i = -1.25$$

On peut écrire :

$$P(x_i \leq 160\,\text{cm}) = P(z \leq -1.25)$$
$$= 0.1056$$

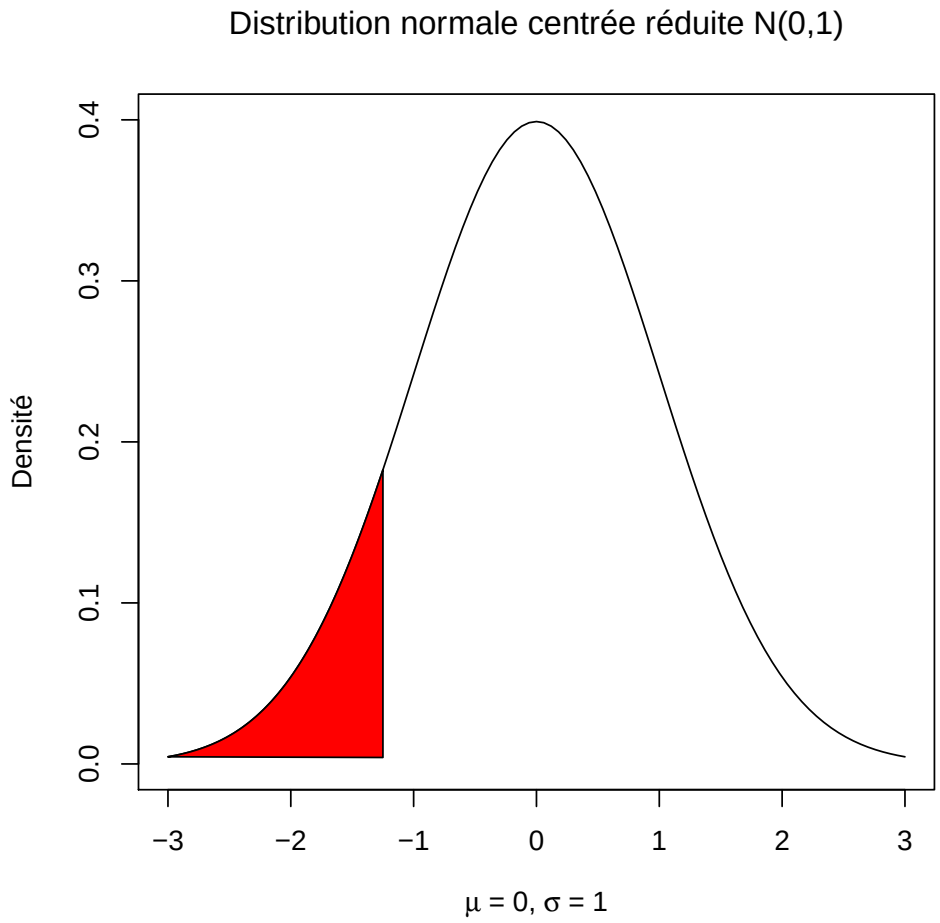

FIGURE 9 – Probabilité cumulative associée à une valeur $\leq$ 160 cm dans une population avec N(170, 8) en se basant sur l'écart normal -1.25 de la valeur 160 cm et sur la distribution normale centrée réduite N(0,1).

Ici, P est une probabilité cumulative que l'on peut obtenir en calculant l'aire sous la courbe pour la portion de la courbe N(0,1) à gauche du point -1.25 (fig. 9).

Exemple 1.11 Si on veut connaître la probabilité qu'un individu ait une taille supérieure à 185 cm dans la même population que celle de notre exemple précédent

(édifice à bureaux), on pourrait encore une fois utiliser l'écart normal. Ainsi, on peut écrire

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{185 - 170}{8}$$
$$z_i = 1.875$$

La probabilité cumulative ici peut s'obtenir à l'aide de :

$$\begin{aligned} P(x_i \leq 185\,\text{cm}) &= P(z \leq 1.875) \\ &= 0.9696 \end{aligned}$$

Toutefois, nous désirons $P(x_i > 185\,\text{cm})$, ce qui diffère des exemples précédents avec $P(x_i \leq X)$ (fig. 10a). L'astuce ici consiste à calculer le complément de $P(x_i \leq 185\,\text{cm})$ (fig. 10b). Puisque l'aire sous la courbe est de 1, on peut obtenir $P(x_i > 185\,\text{cm})$ simplement avec :

$$\begin{aligned} P(x_i > 185\,\text{cm}) &= 1 - P(x_i \leq 185\,\text{cm}) \\ &= 1 - 0.9696 \\ &= 0.0304 \end{aligned}$$

Exemple 1.12 Il est possible de déterminer la probabilité d'observer une valeur entre 165 cm et 180 cm dans la même population. Pour ce faire, il faut calculer

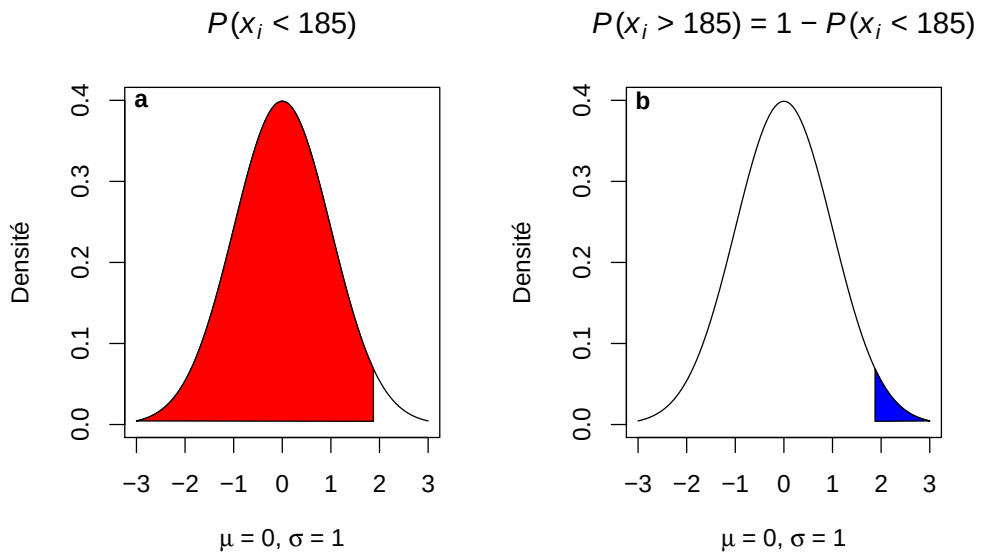

FIGURE 10 – (a) Probabilité cumulative associée à une valeur $<$ 185 cm et (b) complément de cette probabilité (1 - P) pour la valeur $>$ 185 cm dans une population avec N(170, 8) en se basant sur l'écart normal 1.875 de la valeur 185 cm et sur la distribution normale centrée réduite N(0,1).

l'écart normal associé à chacune des bornes, comme suit :

$$z_1 = \frac{180 - 170}{8} = 1.25$$
$$z_2 = \frac{165 - 170}{8} = -0.625$$

On obtient les probabilités cumulatives de chaque z_i comme d'habitude :

$$P(x_1 \leq 180\,\text{cm}) = P(z_1 \leq 1.25) = 0.8944$$

$$P(x_2 \leq 165\,\text{cm}) = P(z_2 \leq -0.625) = 0.2660$$

La différence entre les deux probabilités cumulatives nous donnera la probabilité

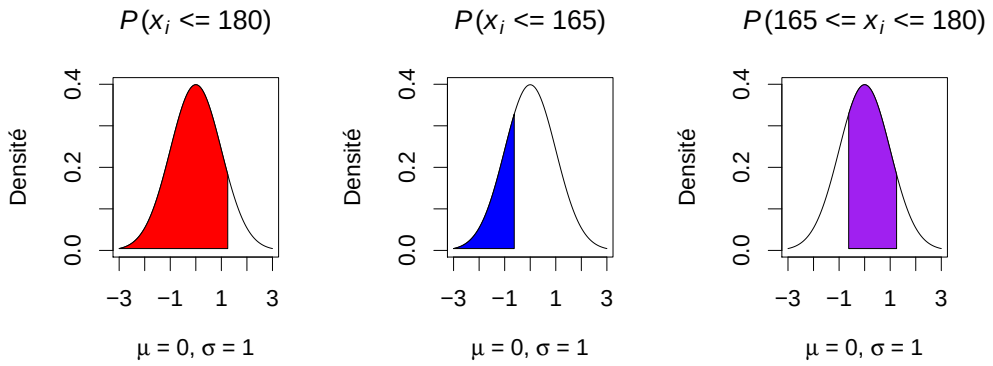

FIGURE 11 – Probabilités cumulatives associées aux valeurs $x_i \leq$180 cm et et $x_i \leq$ 165 cm et probabilité d'obtenir 165 cm $\leq x_i \leq$ 180 cm dans une population N(170,8) en se basant sur les écarts normaux -0.625 et 1.25 pour les valeurs de 165 et 180 cm respectivement, et sur la distribution normale centrée réduite N(0,1).

pour l'intervalle désiré (fig. 11) :

$$\begin{aligned} P(165\,\text{cm} \leq x_i \leq 180\,\text{cm}) &= 0.8944 - 0.2660 \\ &= 0.6284 \end{aligned}$$

7.4 Applications de la statistique z à un échantillon

Comme nous venons de le voir, la statistique z peut s'appliquer aux observations d'une population dont on connaît les réelles valeurs de la moyenne et de l'écart-type. On utilise la distribution normale pour trouver la probabilité d'observer x_i (une valeur d'une observation) dans un intervalle donné d'une population avec une moyenne μ et un écart-type σ connus. Toutefois, on peut aussi appliquer la même approche au niveau d'un paramètre d'une population. En d'autres mots, on peut déterminer l'écart normal (z) associé à la valeur de l'estimation d'un paramètre à partir d'un échantillon (p. ex., une moyenne, une médiane)

d'une population avec μ et σ connus.

Ainsi, l'équation originale $z = \frac{x_i - \mu}{\sigma}$ qui dépendait de la normalité des observations devient $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ et s'intéresse à un paramètre d'une population, où $\bar{x}$ correspond à la moyenne arithmétique de l'échantillon, et $\sigma_{\bar{x}}$ est l'erreur-type de la population. Autrement dit, au lieu de s'intéresser à des valeurs individuelles dans la population, nous ciblons plutôt la moyenne d'un échantillon tiré d'une population avec μ et $\sigma_{\bar{x}}$ connus. Cette transition est possible en supposant que la moyenne provienne d'une distribution normale des moyennes (grâce au théorème de la limite centrale). C'est le genre de traitement typique que nous faisons la plupart du temps avec les données d'un échantillon que nous récoltons à partir d'une population statistique.

Exemple 1.13 Nous voulons déterminer la probabilité d'obtenir un échantillon aléatoire de 9 longueurs de becs d'oiseaux, lequel a une moyenne > 50.0 mm dans une population avec $\mu = 47.5$ mm et $\sigma = 12.89$ mm. On obtient :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12.89}{\sqrt{9}} = 4.30$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{50 - 47.5}{4.30} = 0.58$$

$$P(\bar{x} > 50\,\text{mm}) = P(z > 0.58) = 0.2803$$

On constate qu'il est assez probable ($P(z > 0.58) = 0.2803$) de tirer un échantillon de 9 longueurs de becs d'oiseaux avec des propriétés similaires à celles de l'échantillon original de la population d'intérêt. Par convention, on considère qu'une probabilité $P \leq 0.05$ est faible, bien que ce seuil soit arbitraire et qu'il existe de nouvelles approches qui mettent de côté la subjectivité du choix d'un tel seuil. Nous discuterons plus en détail des implications du choix de tels seuils ainsi que de méthodes alternatives dans les prochaines leçons.

Malheureusement, on connaît rarement σ dans la vraie vie et on doit utiliser une estimation obtenue à partir de l'échantillon. Comme nous l'avons souligné plus tôt, l'estimation de la variance est difficile dans les échantillons de petite taille. Comme le test z nécessite une estimation de la variance, ce dernier est peu utilisé pour tester des valeurs d'un échantillon de petite taille. Lorsque σ est inconnu et doit être estimé à partir d'un échantillon, nous utiliserons plutôt la distribution du t de Student. C'est ce que nous verrons dans les prochaines leçons.

Conclusion

Dans cette leçon, nous avons brièvement présenté les concepts de base importants en statistique, notamment les caractéristiques d'une population et d'un échantillon, les variables aléatoires, les mesures de tendance centrale et de dispersion. La distribution normale a été présentée, ainsi que le théorème de la limite centrale et la loi des grands nombres.

Index