Simplifions la statistique

Par Louis Blais, stat., ASSQ

ERREUR ET NIVEAU DE CONFIANCE : QUE SIGNIFIENT-ILS?

L'information concernant l'erreur autour de la moyenne dans les sondages sur les intentions de vote publiés dans les journaux est souvent libellée comme ceci :

> **« …ce sondage a une marge d'erreur de 3,1 %, 19 fois sur 20. »**

Cela signifie-t-il que la marge d'erreur n'est pas de 3,1 %, une fois sur 20? La réponse est non. La réalité est beaucoup plus subtile et complexe que cela. Cet exemple, tiré de la presse écrite, vaut également en foresterie. La marge d'erreur peut s'appliquer sur un volume moyen ou sur une surface terrière moyenne, tout comme le « 19 fois sur 20 ». Mais que signifient-ils en fait?

*Pour éviter toute confusion, l'erreur autour de la moyenne sera nommée « erreur » dans le présent texte.

L'objectif principal d'un sondage est de connaître la vraie valeur de la population étudiée. Comme ce nombre est quasi impossible à obtenir pour des raisons budgétaires, un échantillonnage est nécessaire. Cet échantillonnage comportera donc une erreur avec laquelle un intervalle de confiance peut être construit.

La borne inférieure de l'intervalle représente la moyenne, moins l'erreur, et la borne supérieure, la moyenne, plus l'erreur. L'erreur est donc toujours liée à la moyenne et l'intervalle obtenu contiendra probablement la vraie valeur de la population.

Forêts, Faune et Parcs
Québec

L'ERREUR

Dans un sondage, peu importe le domaine d'application, l'erreur dépend des valeurs obtenues de chaque unité d'échantillonnage. Supposons par exemple que la population étudiée est un territoire qui contient une quantité précise de bois marchand, exprimée en volume par hectare. Si un échantillon de 50 placettes, sur un territoire de 200 ha, a obtenu un volume moyen de 150 m³/ha, un autre échantillon de 50 placettes disposées à des endroits différents du même territoire aura, fort probablement, une autre moyenne, par exemple, de 160 m³/ha. De plus, il y a fort à parier que l'erreur obtenue dans les deux sondages sera différente. Si l'erreur relative est la même, disons 10 %, l'erreur obtenue dans le premier sondage sera de 15 m³/ha alors qu'elle sera de 16 m³/ha dans le second. Pire, il n'est pas garanti que l'erreur relative soit la même dans deux sondages différents portant sur la même population. Voyons pourquoi.

L'erreur dépend principalement de quatre éléments :

- **L'écart-type**. C'est, à peu de choses près, l'écart moyen des observations par rapport à la moyenne des échantillons. C'est une mesure de la variabilité des observations autour de la moyenne. Plus l'étendue des observations autour de la moyenne augmente, plus l'écart-type augmente. Si l'écart-type double, et que tous les autres éléments sont égaux, l'erreur sera double également.
- **La valeur de *t* de Student.** Elle est fonction d'une certaine probabilité et du nombre d'unités d'échantillonnage. Plus la valeur de *t* de Student est petite, plus l'erreur diminue. Cette valeur peut être très élevée lorsque le nombre d'unités d'échantillonnage est égal à 2 et diminue rapidement pour se stabiliser à partir de 30 unités d'échantillonnage. Le tableau 1 présente les valeurs de *t* de Student en fonction du nombre d'unités d'échantillonnage pour un niveau de confiance à 95 %.

TABLEAU 1 Valeur de *t* de Student pour un niveau de confiance à 95 % en fonction du nombre d'unités d'échantillonnage (n)

n	*t* de Student
2	12,706
3	4,303
4	3,182
5	2,776
10	2,262
15	2,145
20	2,093
25	2,064
30	2,045
35	2,032
40	2,023
45	2,015
50	2,010

- **Le nombre d'unités d'échantillonnage.** Il contribue à faire diminuer l'erreur. Il est en effet logique que plus le nombre d'unités d'échantillonnage augmente, plus l'erreur diminue. Le nombre d'unités d'échantillonnage compense pour un écart-type plus grand, puisque, dans le calcul, il est au dénominateur, contrairement à l'écart-type qui est au numérateur. Il influence également la valeur de *t* de Student. Il contribue à faire diminuer cette valeur lorsque le nombre augmente et à faire diminuer l'erreur.

- **Le niveau de confiance.** Il influence la valeur de *t* de Student. Plus le niveau de confiance est élevé, plus la valeur de *t* de Student sera élevée et fera augmenter l'erreur. Le tableau 2 présente la valeur de *t* de Student en fonction du niveau de confiance pour 50 unités d'échantillonnage.

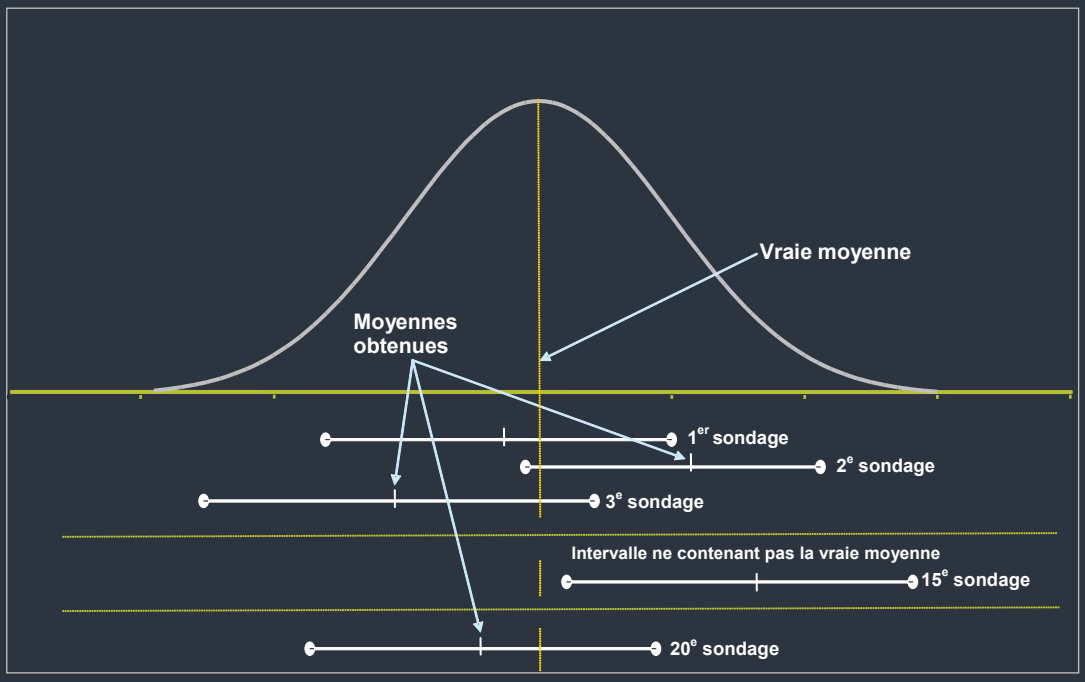
TABLEAU 2 Valeur de *t* de Student pour 50 unités d'échantillonnage en fonction du niveau de confiance

Niveau de confiance (%)	*t* de Student
85	1,462
90	1,677
92	1,788
94	1,925
95	2,010
99	2,680

En pratique, obtenir une erreur acceptable est un compromis entre le nombre d'unités d'échantillonnage (coûts) et le niveau de confiance désiré (risque), deux éléments que l'on doit déterminer avant de procéder au sondage. Il est possible d'obtenir une estimation de l'écart-type à partir de sondages précédents pour connaître l'étendue des valeurs auxquelles il faut s'attendre.

LE NIVEAU DE CONFIANCE

Le niveau de confiance est relié à une probabilité. C'est l'élément de risque du sondage. Il faut être conscient qu'on ne peut jamais garantir à 100 % que la vraie moyenne d'un sondage se trouve dans l'intervalle obtenu. C'est une question de hasard. Or, qui dit hasard, dit probabilité. S'il était possible de faire un très grand nombre de sondages sur un même territoire, un certain nombre d'entre eux contiendrait la vraie moyenne. Le niveau de confiance représente le pourcentage des sondages dont l'intervalle contiendrait la vraie moyenne. Ce pourcentage est le niveau de confiance. Un niveau de confiance à 95 % représente la probabilité que 19 fois sur 20 ou que dans 19 intervalles sur 20 on obtienne la vraie moyenne. La figure suivante illustre ce propos.



Il est possible d'augmenter le niveau de confiance, mais alors l'erreur augmentera aussi. Pour compenser l'augmentation de l'erreur occasionnée par la hausse du niveau de confiance, il faut augmenter le nombre d'unités d'échantillonnage, ce qui augmente les coûts.

En conclusion, pour reprendre le libellé classique des journaux, la marge d'erreur est l'erreur autour de la moyenne obtenue, et le « 19 fois sur 20 » est l'élément de risque du sondage, c'est-à-dire qu'une fois sur 20, l'intervalle autour de la moyenne ne contiendra pas la vraie moyenne. C'est ce qu'il faut comprendre de ce libellé.

POUR EN SAVOIR PLUS...

Ministère des Forêts, de la Faune et des Parcs
Direction de l'aménagement et de l'environnement forestiers
5700, 4e Avenue Ouest
Québec (Québec) G1H 6R1
daef@mffp.gouv.qc.ca

N° de publication : DAEF2-0069-F-4

MFFP - Octobre 2014