Simplifions la statistique

Par Louis Blais, stat., ASSQ

TEST DE DIFFÉRENCE ENTRE DEUX MOYENNES

Le test de différence entre deux moyennes est généralement utilisé en foresterie lorsqu'on est en présence de deux inventaires parallèles sur le même territoire. C'est l'une des méthodes utilisées pour vérifier un inventaire réalisé par un entrepreneur. Ce test est bien documenté dans la littérature, mais plusieurs affirmations ont été véhiculées sur la façon de le faire, particulièrement par intervalles de confiance.

Le but de cette fiche est de démontrer la fausseté de certaines affirmations et de montrer la façon correcte de faire le test au moyen d'un intervalle de confiance.

Dans le passé, il y a eu plusieurs affirmations sur la façon de faire ce test par intervalles de confiance. Voici deux affirmations qui sont particulièrement répandues.

- **Il n'y a pas de différence significative entre les moyennes si les intervalles de confiance des deux sondages se croisent. Autrement, c'est différent.**

- **Il y a une différence significative si la moyenne du premier sondage n'est pas dans l'intervalle de confiance du deuxième sondage, et inversement.**

Forêts, Faune et Parcs
Québec

Les deux affirmations précédentes sont fausses. Voyons comment construire correctement un intervalle de confiance à une différence entre deux moyennes, puis comment faire le test approprié.

CONSTRUCTION D'UN INTERVALLE DE CONFIANCE : CAS GÉNÉRAL

On construit l'intervalle de confiance à partir de l'erreur du sondage. Le calcul des bornes se fait à l'aide de la formule générale suivante :

$$[\bar{x} - erreur\ ,\ \bar{x} + erreur]$$

où l'erreur est calculée comme suit :

$$\text{erreur} = (\text{t de Student})\,(\text{erreur type})$$
$$= (t_{(n-1,1-\alpha/2)}) \quad \left(\sqrt{s^2/n}\right)$$

Dans cette formule :

- $\bar{x}$ est la moyenne;
- t est la valeur de Student;
- s^2 est la variance de l'échantillon;
- n est le nombre d'unités d'échantillonnage;
- (1-α) % est le niveau de confiance de l'intervalle.

Donc, si les valeurs de chaque unité d'échantillonnage sont connues, il est possible de calculer la moyenne et la variance. Puis, en utilisant le nombre d'unités d'échantillonnage et la valeur appropriée de *t* de Student, on peut alors calculer l'intervalle de confiance.

Le deuxième terme de la formule est l'erreur type. Elle est établie en fonction de l'écart-type des valeurs obtenues dans le sondage et du nombre d'unités d'échantillonnage. Plus il y a d'unités d'échantillonnage, plus l'erreur type diminue.

CONSTRUCTION D'UN INTERVALLE DE CONFIANCE : CAS D'UNE DIFFÉRENCE ENTRE DEUX MOYENNES

Il faut savoir que chaque statistique utilisée a une marge d'erreur construite avec sa variance. C'est évidemment le cas pour une moyenne, mais ça l'est également pour une différence de moyenne. Le changement avec le cas général se situe au niveau de la moyenne ($\bar{x}$) et du calcul de l'erreur type. Comme l'intérêt est à la différence entre deux moyennes, il faut remplacer la moyenne par la différence entre les deux moyennes ($\bar{x}_1 - \bar{x}_2$) et l'erreur type par l'erreur type de la différence. En clair, il faut remplacer la formule du cas général dans le calcul de l'erreur par celle-ci (dans le cas où les variances sont égales) :

$$erreur_{diff} = (\text{t de Student})\,(\text{erreur type}_{diff})$$
$$= (t_{(n_1+n_2-2,\,1-\alpha/2)}) \left(\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}\right)$$

Dans cette formule :

- l'indice « 1 » se rapporte aux données du sondage 1;
- l'indice « 2 » se rapporte aux données du sondage 2.

Chaque jeu de données a sa variance (et) et son nombre d'unités d'échantillonnage (et).

Noter le nombre d'unités d'échantillonnage total à utiliser pour trouver la valeur correspondante du *t* de Student. Il faut noter également que si l'un des deux sondages a plus d'unités d'échantillonnage que l'autre, il aura plus de poids dans le calcul. Cela correspond donc à une somme pondérée des deux variances des sondages.

Puis, il faut calculer les bornes de l'intervalle de confiance comme suit :

$$\left[(\bar{x}_1 - \bar{x}_2) - erreur_{diff}\,,\,(\bar{x}_1 - \bar{x}_2) + erreur_{diff}\right]$$

TEST DE LA DIFFÉRENCE ENTRE DEUX MOYENNES AVEC UN INTERVALLE DE CONFIANCE DE LA DIFFÉRENCE

Une fois que l'intervalle de confiance de la différence est construit, il suffit de vérifier si la valeur nulle ou zéro est incluse dans cet intervalle pour faire le test correctement.

En clair :

- Si la borne inférieure et la borne supérieure sont négatives, la valeur nulle ou « 0 » n'est pas incluse dans l'intervalle, alors les deux moyennes sont significativement différentes et la moyenne du sondage 2 est supérieure à celle du sondage 1.

- Si la borne inférieure et la borne supérieure sont positives, alors les deux moyennes sont significativement différentes et la moyenne du sondage 1 est supérieure à celle du sondage 2.

- Si la borne inférieure est négative et la borne supérieure est positive, la valeur nulle ou « 0 » est forcément à l'intérieur de l'intervalle de confiance et les deux moyennes ne sont pas significativement différentes.

Voici deux exemples avec un niveau de confiance de 95 %.

Exemple 1. La moyenne du sondage 1 n'est pas dans l'intervalle de confiance du sondage 2 et inversement.

Le test sur la différence entre les deux moyennes montre que la différence n'est pas significative alors que, selon l'affirmation véhiculée, la conclusion aurait dû prouver le contraire.

Sondage	n	Erreur type	Moyenne	Erreur	Borne inférieure	Borne supérieure
1	48	1,443	50	2,90	47,10	52,90
2	50	1,414	54	2,84	51,16	56,84
Différence	96	2,021	4	4,01	-0,11	8,011

Exemple 2. Les deux intervalles de confiance se chevauchent. Le test sur la différence entre les deux moyennes montre que la différence est significative alors que, selon l'affirmation véhiculée, la conclusion aurait dû prouver le contraire.

Sondage	n	Erreur type	Moyenne	Erreur	Borne inférieure	Borne supérieure
1	48	2,021	50	4,06	45,94	54,06
2	50	1,414	55	2,84	52,16	57,84
Différence	96	2,450	5	4,86	0,137	9,863

Il est important de remarquer que le calcul de la variance de chaque sondage n'est pas précisé. Ce calcul dépend de la méthode d'échantillonnage. Par exemple, le calcul de la variance n'est pas le même pour un échantillonnage systématique que pour un échantillonnage stratifié. Il existe plusieurs volumes sur le sujet et il vaut mieux s'y référer.

En terminant, il importe de retenir que le croisement ou non des intervalles de chaque sondage ou le fait que la moyenne ne soit pas dans l'intervalle de l'autre, ou inversement, ne sont pas de bons moyens pour conclure à la présence ou à l'absence d'une différence des deux moyennes. Pour y arriver, il faut utiliser le test de la différence et les statistiques appropriées.

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

Document accessible sur le site intranet :
http://www.intranet/forets/outils-documentation/simple_statistique.asp

N° de publication : DAEF2-0069-F-2

MFFP - Octobre 2014