Méthodes de sondage – SOL3017 et SOL6448

Notes de cours

Dix-neuf fois sur vingt, la marge d'erreur

Département de sociologie
Université de Montréal

Professeur : Claire Durand

1. Marge d'erreur et taille requise de l'échantillon

1.1 Les distributions, la distribution d'échantillonnage

Il y a trois types de distributions:

Deux sont bien connues:

- La distribution d'un échantillon à laquelle sont associées des statistiques

moyenne de l'échantillon : $\bar{X}$
écart-type de l'échantillon : s
variance de l'échantillon : s^2

- La distribution de la population à laquelle sont associés des paramètres

moyenne de la population : μ
écart-type de la population : σ
variance de la population : σ^2

Si on prenait tous les échantillons que l'on peut tirer d'une même population, on obtiendrait une distribution des caractéristiques des échantillons, distribution à laquelle sont associées les statistiques de la distribution d'échantillonnage:

La moyenne des échantillons : Elle s'approche de la moyenne de la population avec l'augmentation du nombre d'échantillons.

L'erreur-type ($s/\sqrt{n}$) est l'écart-type de la distribution des moyennes des échantillons.

Pour une proportion, l'erreur-type= $\frac{s}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$

Quelque soit le type de distribution de la population, la distribution des moyennes des échantillons tirés de la population tendra vers une distribution normale avec l'augmentation du nombre d'échantillons tirés. Cette distribution aura une moyenne μ et une variance de σ^2/n. C'est le théorème central limite.

CECI CONSTITUE LA BASE DE TOUTES LES STATISTIQUES INFÉRENTIELLES.

1.2 Le calcul de la marge d'erreur en fonction du seuil de confiance recherché

- SEUIL DE CONFIANCE α

La probabilité qu'un échantillon représente bien une population, étant donné les lois des probabilités, se nomme le ***seuil de confiance.*** *C'est la certitude que l'on a quant à la justesse des résultats*. Le critère que l'on retient habituellement est de 95%, **c'est-à-dire que si on prend plusieurs échantillons d'une même population, *19 fois sur 20* (95% des fois), l'échantillon constituera une représentation fiable de cette population, soit à l'intérieur de la marge d'erreur. Cette proportion correspond à 1,96 écart-type sur la courbe normale.** Cette valeur est le Z_{α}, c'est-à-dire la surface sous la courbe normale pour 1-α (i.e. 0,95).

Si l'on désirait un seuil de confiance plus grand, 99% par exemple, cette proportion correspondrait à un Z_{α} de 2,58 alors qu'à l'inverse, un seuil de confiance de 90% correspond à un Z_{α} 1,67.

- MARGE D'ERREUR - INTERVALLE DE CONFIANCE :

La marge d'erreur, *c'est* ***la précision du résultat obtenu étant donné le seuil de confiance que l'on est prêt à accepter.*** La ***marge d'erreur*** est alors égale à Z_{α} **erreur-type* et la formule est la suivante pour une proportion:

$$e = Z\alpha * \sqrt{\frac{p(1-p)}{n}}$$

où Z_{α} est la surface où l'on retrouve 1-α de la courbe normale (Z_{α}) et donc 1,96 lorsque le seuil de confiance accepté est de 95%,
p est la proportion de personnes ayant le comportement dont on estime la précision,
et *n* est la taille de l'échantillon.

Pour un seuil de confiance de 95%, lorsque la proportion est de 30% (0,30), cette formule donne:

$$e = 1{,}96 * \sqrt{\frac{0{,}30(1-0{,}30)}{800}} = 1{,}96 * \sqrt{\frac{0{,}21}{800}} = 1{,}96 * \sqrt{(0{,}0002625} = 1{,}96 * {,}0162 = 0{,}031752$$

Et donc, si on veut transformer en pourcentage:

e%=e*100= 3,1752

Lorsque la proportion est maximale (0,5), cette formule peut être simplifiée:

$$e\% = 1{,}96 * \sqrt{\frac{0{,}50(1-0{,}50)}{n}} * 100 = 1{,}96 * \frac{\sqrt{0{,}25}}{\sqrt{n}} = 1{,}96 * (0{,}5) * \frac{1}{\sqrt{n}} \approx \frac{1}{\sqrt{n}}$$

Elle est donc presque équivalente à $1/\sqrt{n}$.

→ D'où la note dans la présentation des résultats de sondage tirés d'échantillon d'environ 1000 personnes:

"La marge d'erreur maximale est de **3,2%**" , soit

$$e\% = 1{,}96 * \sqrt{\frac{,5(1-,5)}{1000}} * 100 \approx \frac{1}{\sqrt{1000}} * 100 = 3{,}16\%$$

et ce **19 fois sur 20** (dans 95% des cas)

Note 1: On calcule habituellement pour les sondages la marge d'erreur maximale, c'est-à-dire celle qui correspond à une proportion de 0,5 (50%). C'est en effet quand la population se divise moitié moitié que la marge d'erreur des résultats est la plus grande. Par contre, pour chacun des résultats présentés, il est possible de calculer la marge d'erreur spécifique. Pour ce faire, il suffit d'utiliser la formule présentée plus haut.

Note 2: Il y a une relation entre la certitude (le seuil de confiance) et la marge d'erreur. On peut constater que plus on veut une grande certitude, plus la marge d'erreur est grande et moins l'estimation est précise; par contre, moins la certitude est grande, plus l'estimation est précise. Par exemple, pour une proportion de 50% et un échantillon de 1000 personnes, je peux être certain à 95% que la proportion se situe entre 46,8% et 53,2% (marge d'erreur de 3,2%), mais je peux être certain à 99% qu'elle se situe entre 46% et 54% (l'erreur standard est alors multipliée par 2,58 plutôt que 1,96 ce qui donne une marge d'erreur de 4%).

Correction pour population finie

Lorsque la population est "finie", c'est à dire que la population est petite soit moins grande que 20 fois l'échantillon, on ajoute ce qu'on appelle une ***correction pour population finie***. La formule de la marge d'erreur devient la suivante:

où Z_α est la surface où l'on retrouve 1-α de la courbe normale (Z_α) et donc 1,96 lorsque le

$$e = Z\alpha * \sqrt{\frac{p(1-p)}{n}} * \sqrt{\frac{(N-n)}{(N-1)}}$$

seuil de confiance accepté est de 95%,
p est la proportion de personnes ayant le comportement dont on estime la précision,
n est la taille de l'échantillon,
et *N* est la taille de la base échantillonnale.

Ainsi, pour un échantillon de 1000 personnes et une base de sondage comprenant 6,800 noms, un seuil de confiance de 95% et une proportion maximale (0,5), on aurait l'équation suivante:

ce qui donne 2,86% (plutôt que 3,16% s'il s'agissait d'un échantillon provenant d'une

$$e = 1{,}96 * \sqrt{\frac{0{,}5(1-0{,}5)}{1000}} * \sqrt{\frac{(6{,}800-1000)}{(6{,}800-1)}}$$

population *infinie*).

L'**intervalle de confiance** d'une proportion est égal à cette proportion ± la marge d'erreur. Ainsi dans l'exemple précédent, la marge d'erreur est de 50% ± 2,86%, c'est-à-dire que la proportion ***dans la population*** se situe vraisemblablement entre 47,14% et 52,86%. C'est ce qu'on appelle l'intervalle de confiance.

Si l'on veut savoir si deux proportions d'un même échantillon sont différentes, une manière simple et précise, mais conservatrice, de le vérifier consiste à calculer les intervalles de confiance des deux proportions et de voir si ces intervalles se chevauchent (le maximum d'un intervalle est plus élevé que le minimum de l'autre). Quelques petits exercices sur le site vous permettent de vérifier si vous avez bien compris ces notions.

De meilleurs tests existent pour vérifier si deux proportions sont significativement différentes. Pour deux proportions d'un même échantillon, la formule est la suivante:

$$ediff = Z\alpha * \sqrt{\frac{(p_1 + p_2) + (p_1 - p_2)_2}{n}}$$

Si l'on veut savoir si deux proportions de deux échantillons différents sont significativement différentes, la formule de la marge d'erreur de la différence est la suivante:

$$e_{diff} = Z_\alpha * \sqrt{\frac{p_1*(1-p_1)}{n_1} + \frac{p_2*(1-p_2)}{n_2}}$$

Où p_1 et p_2 sont les proportions respectives dont on calcule la différence et n_1 et n_2, les tailles des échantillons. Cette formule peut être simplifiée par la formule suivante qui donne des résultats très proches:

$$e_{diff} = Z_\alpha * \sqrt{\frac{2\,p*(1-p)}{n}}$$

Où p est la moyenne pondérée de p_1 et p_2 et n, la moyenne de n_1 et n_2. Ces formules s'approchent de 1,4 fois la marge d'erreur de la proportion. Notez que cette formule n'est pas matière à examen.

Note: On peut trouver sur le site web un tableau qui donne les résultats approximatifs de ces calculs pour diverses proportions et marges d'erreur avec un seuil de confiance de 95%. On peut également consulter différents sites sur Internet (dont Circum http://circum.com/index.cgi?fr:doc) qui donnent le résultat de ces calculs. Il suffit alors de savoir quel chiffre mettre à quel endroit!