> La taille d'échantillon

La taille de l'échantillon est un facteur primordial pour obtenir des résultats fiables. Elle va varier en fonction de la précision souhaitée sur l'ensemble de l'échantillon et sur les sous-cibles étudiées.

La précision peut se mesurer par l'intervalle de confiance ou marge d'erreur statistique des estimations. Plus on recherche un intervalle de confiance petit, plus la taille de l'échantillon devra être importante.

Estimation d'une proportion

Pour déterminer la taille de l'échantillon avec une précision d, la formule statistique pour un tirage aléatoire simple est :

$$n = \frac{(z)^2 p\,(1-p)}{d^2}$$

n = taille de l'échantillon
z = niveau de confiance selon la loi normale centrée réduite :
- pour un niveau de confiance de 95%[1], z = 1.96,
- pour un niveau de confiance de 99%, z = 2.575)

p = proportion estimée de la population qui présente la caractéristique
d = marge d'erreur tolérée (par exemple on veut connaître la proportion réelle à +/- 5%

Lorsque p = 50% (pour une variable suivant la loi binomiale, l'intervalle de confiance est le plus large lorsque la valeur mesurée est de 50%) $n = \frac{(z)^2}{4d^2}$

Exemples

1) Pour calculer une proportion avec un niveau de confiance de 95% et une marge d'erreur à 5% nous obtenons donc $n = \frac{(1{,}96)^2}{4(0{,}05)^2}$ = 384

2) Pour calculer une proportion avec un niveau de confiance de 99% et une marge d'erreur à 2% nous obtenons donc $n = \frac{(2{,}575)^2}{4(0{,}02)^2}$ = 4144

Les échantillons standards sont souvent de l'ordre de 1000 personnes, notamment dans le cadre des enquêtes Omnibus ou des sondages politiques : c'est un juste milieu entre l'exigence de précision et la faisabilité pratique de l'enquête. Dans le cas de deux réponses possibles à une question (loi binomiale), réparties également (50% de « oui » et de « non » par exemple), la marge d'erreur est de +/- 3% avec une probabilité de 95%.

Cette formule est fréquemment utilisée aussi pour un échantillon par quotas sous l'hypothèse que sa marge d'erreur statistique ne dépasse pas celle du tirage aléatoire simple.

[1] Un niveau de confiance de 95% signifie que l'intervalle de confiance autour de la valeur observée a 95 chances sur 100 de contenir la valeur réelle.

Estimation d'une proportion sur une sous-population

Pour calculer la taille de l'échantillon de manière à estimer une proportion pour une sous-partie de la population, la formule doit être modifiée en

$$n = \frac{(z)^2 p\,(1-p)}{Q\,d^2}$$

avec Q : proportion d'individus appartenant à la sous-population.
Supposons que l'on cherche à mesurer une intention de vote parmi les 18-24 ans (Q=10.7 %). Avec un niveau de confiance de 95 % et une marge d'erreur de 5 %, on obtient en se plaçant dans le pire des cas (p = 50%) :

$$n = \frac{(1{,}96)^2}{4 * 0{,}107 * (0{,}05)^2} = 3\,590$$

La raison est simplement qu'il faut tirer 3 590 individus pour avoir à l'intérieur environ 384 individus âgés de 18 à 24 ans.

Estimation d'une moyenne m

La marge d'erreur de l'estimateur de la moyenne m est $\pm z \frac{\sigma}{\sqrt{n}}$ où σ est l'écart-type de la variable quantitative étudiée, inconnu a priori.

La marge d'erreur relative vaut $\pm z \frac{\sigma}{m\sqrt{n}} = \pm z \frac{CV}{\sqrt{n}}$ où *CV* est le coefficient de variation $\frac{\sigma}{m}$.

Exemple :
Si *CV* = 1, ce qui correspond au cas pas très favorable d'une variabilité de l'ordre de la moyenne, on trouve pour n=1 000 et un niveau de confiance de 95% une marge d'erreur relative de $\pm z \frac{CV}{\sqrt{n}} = \pm \frac{1.96}{\sqrt{1000}} = \pm 6\%$

Ce type de démarche peut également s'appliquer dans des situations plus complexes, comme par exemple :

- Définir un **échantillonnage par grappes :** Pour réaliser un échantillonnage par grappes, il faut partitionner la population mère en grappes regroupant les individus à enquêter, réaliser un tirage aléatoire parmi les grappes (tirage aléatoire simple, méthode des quotas …), puis interroger l'ensemble des individus des grappes sélectionnées (ex : l'ensemble des individus des foyers sélectionnés, l'ensemble des établissements des entreprises sélectionnées, l'ensemble des habitants des Iris sélectionnés dans une commune). La qualité de l'estimation de la mesure sur l'échantillon total dépend du nombre de grappes et de la similarité entre les unités d'une même grappe ;
- Estimer **une différence significative entre 2 proportions**, que ce soit pour la mesure de 2 critères distincts sur une même population ou pour la mesure d'un critère sur 2 sous populations,
- Réaliser un sondage sur une population mère de taille restreinte, et donc présentant un taux de sondage important.

Dans ces cas plus complexes, il convient de se référer aux statisticiens experts des méthodes d'échantillonnage dans les instituts de sondage.