Cours Méthodes de sondage

Dix-neuf fois sur 20, échantillon, marge d'erreur, intervalle de confiance

© Claire Durand, 2024
Département de sociologie, Université de Montréal

Les distributions...

Trois types de distribution, deux bien connus...

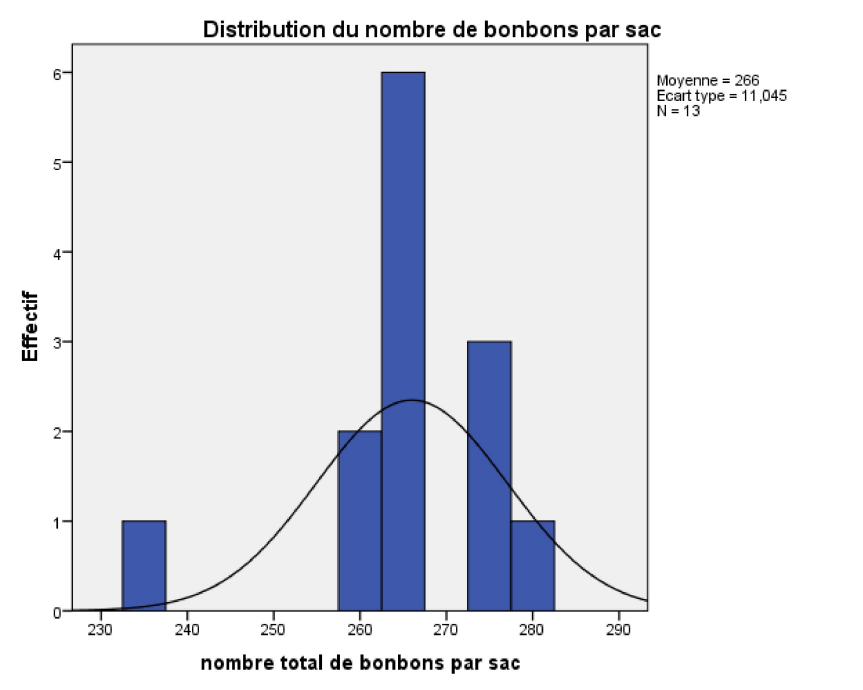
- Distribution de la population:
  - moyenne de la population: μ
  - écart-type de la population: σ
  - variance de la population: σ^2
- Distribution de l'échantillon :
  - moyenne de l'échantillon: $\overline{X}$
  - écart-type de l'échantillon: s
  - variance de l'échantillon: s^2
- Souvent, la distribution du paramètre dans la population (intention de vote pour un parti, proportion de victimes de racisme, etc.) n'est pas connu. C'est ce que nous voulons estimer.

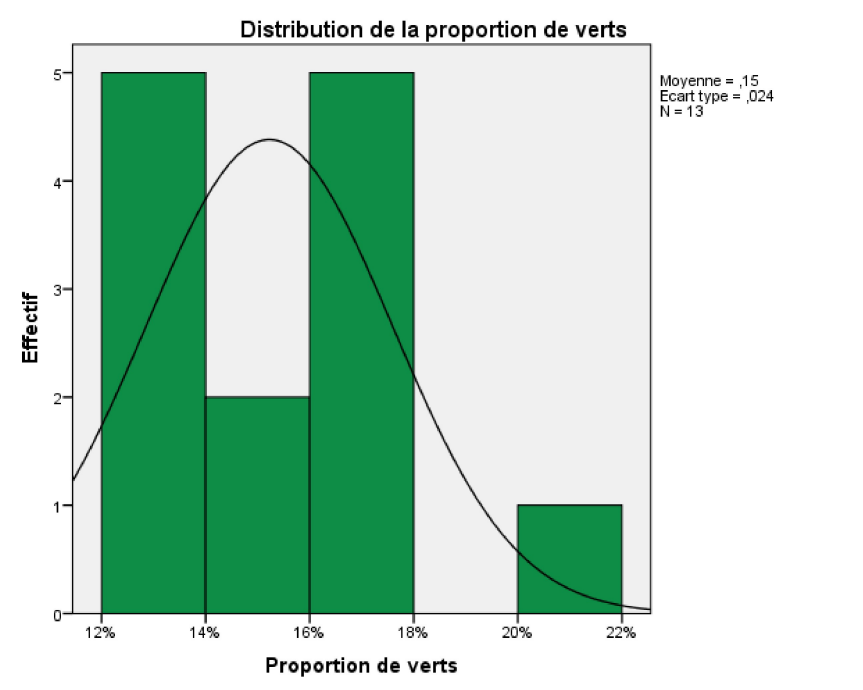
Les distributions

Le ***troisième type***... la distribution des échantillons

- Si on prenait tous les échantillons que l'on peut tirer d'une même population, et que pour chacun, on prenait la moyenne -- ou la proportion -- de l'information qui nous intéresse, on obtiendrait la distribution de la moyenne -- ou la proportion -- d'une caractéristique donnée pour l'ensemble des échantillons.
- Théorème central limite :
  - **Avec l'augmentation du nombre d'échantillons, la moyenne d'une caractéristique des échantillons s'approchera de la moyenne du paramètre de la population.**

13 échantillons aléatoires de 200 grammes de M&M, ça donne quoi?





Le théorème central limite

- ***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 éventuellement une moyenne égale à celle de la population (μ) et une variance de σ^2/n .***
  - ***Cette distribution suit la loi normale (courbe de Gauss)***

Résumé des “notions”

- ℓ : marge d’erreur
- ℓ_{max} : marge d’erreur maximale quand p=0,5 (50%)
- Z_α: Valeur de Z pour un seuil de confiance 1-α
- p: proportion de présence d’une caractéristique dans l’échantillon
- n: taille de l’échantillon
- N: taille de la base échantillonnale

Le seuil de confiance...

- C’est la probabilité qu'un échantillon représente bien une population. *C'est la certitude que l'on veut (peut) avoir 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 fidèle (à l’intérieur de la marge d’erreur) de cette population.**
- **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 (soit 95% de la courbe).

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 égale à Z_α *multiplié par l'erreur-type*.
- La formule est la suivante: $$e = Z_\alpha * \sqrt{\frac{p^*(1-p)}{n}}$$
  - Où :
    - Z_α est l'espace sous la courbe normale (1,96 quand α = 0,05; ça serait 2,58 pour α = 0,01)
    - ***p* est la proportion que l'on retrouve dans l'échantillon et dont on cherche la marge d'erreur**
    - *1-p* est 1 moins la proportion
    - ***n* est la taille de l'échantillon *sur lequel* porte la proportion**.

Concrètement,

Pour une proportion de 30% (p=0,30) d'un échantillon comprenant 1000 unités/répondants

$$e = 1{,}96 * \sqrt{\frac{0{,}3*(1-0{,}3)}{1000}} = 1{,}96 * \sqrt{\frac{0{,}21}{1000}} = 1{,}96 * \sqrt{0{,}00021} = 1{,}96 * 0{,}01449$$

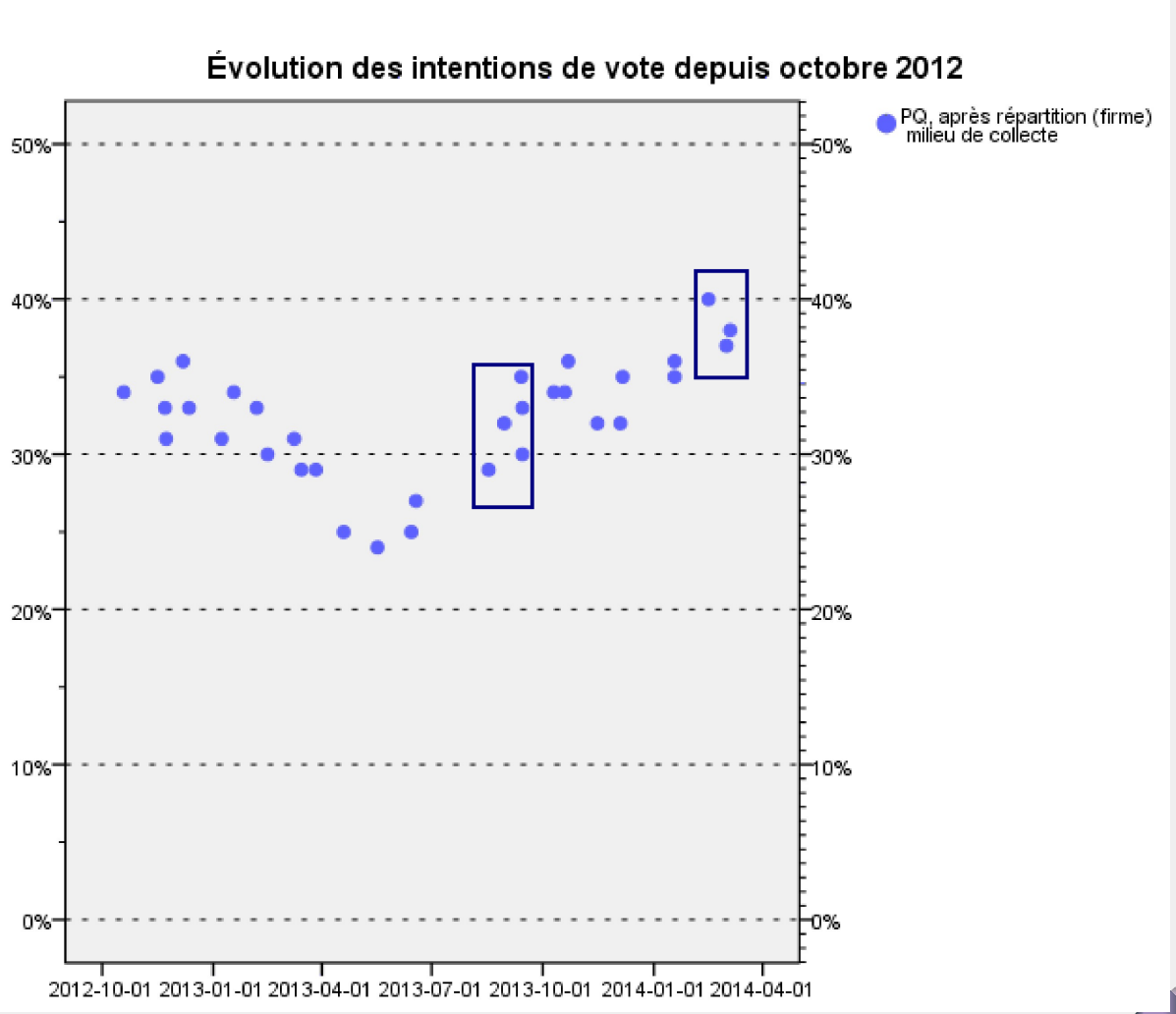
$$e = 0{,}0284 \quad et \quad e\% = e*100 = 2{,}84\%$$

- ET donc 19 fois sur 20 (dans 19 échantillons sur 20, *soit 95% des échantillons*), la proportion se situera entre 30% - 2,84% et 30% + 2,84%, soit entre 27,16% et 32,84%. C'est ***l'intervalle de confiance*** de la proportion.

Exemple d’impact de la proportion et de la taille

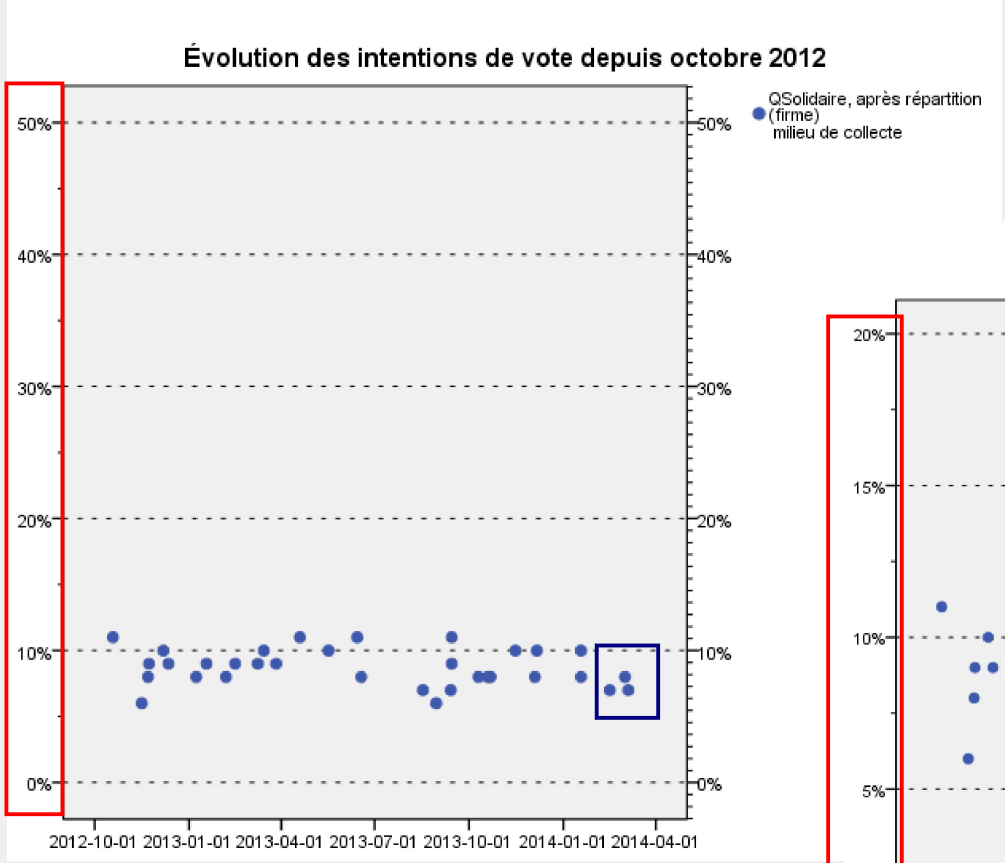
Intentions de vote PQ en 2014

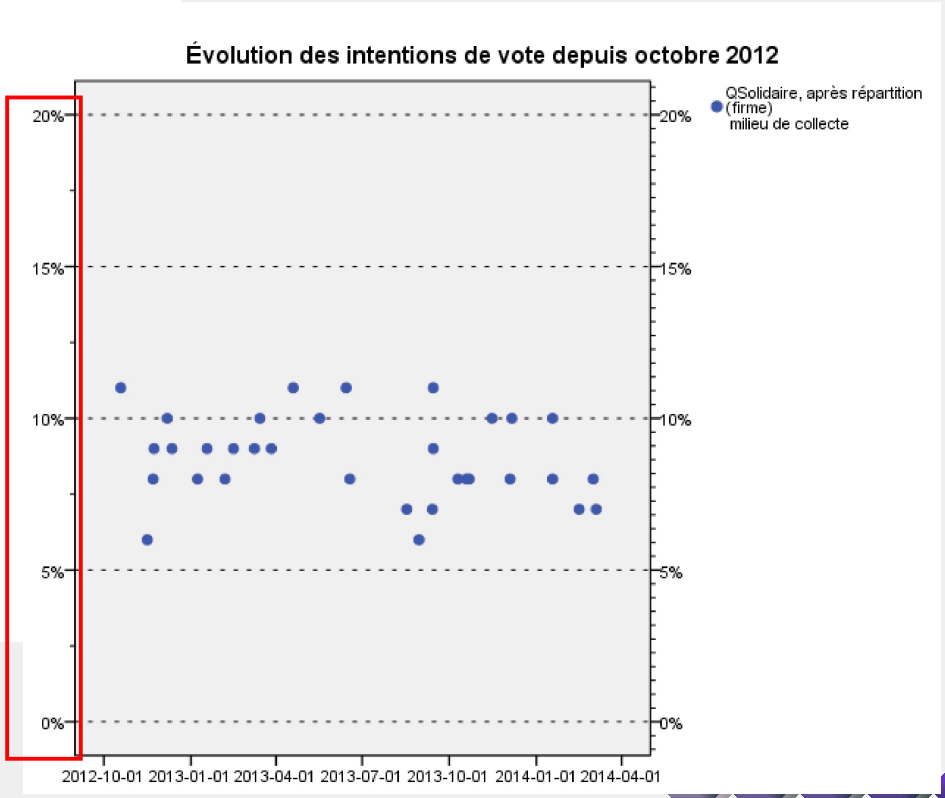
L'effet des proportions



Intentions de vote QS en 2014

Effet de la proportion (et des échelles)

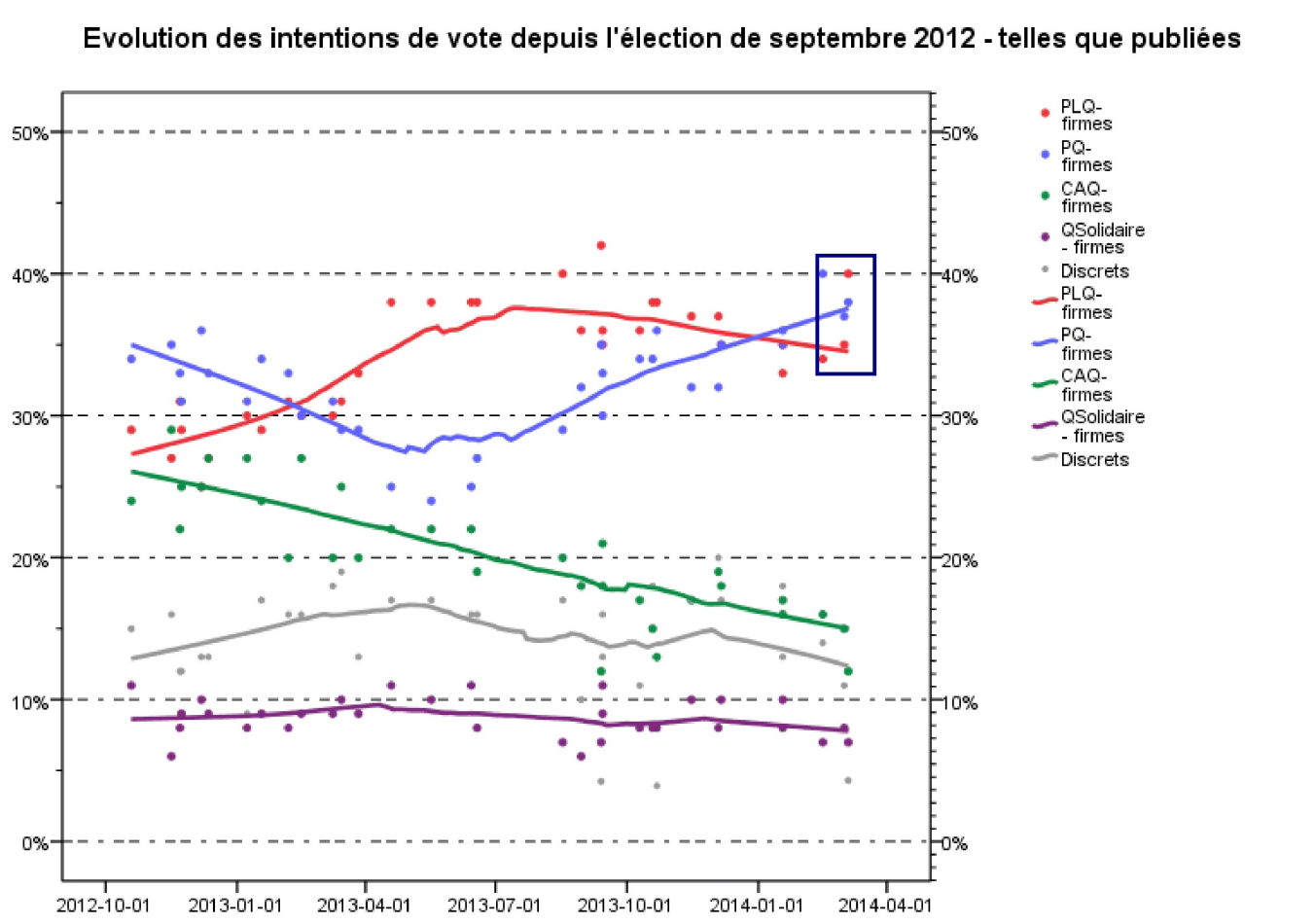




L'effet des proportions et des échelles

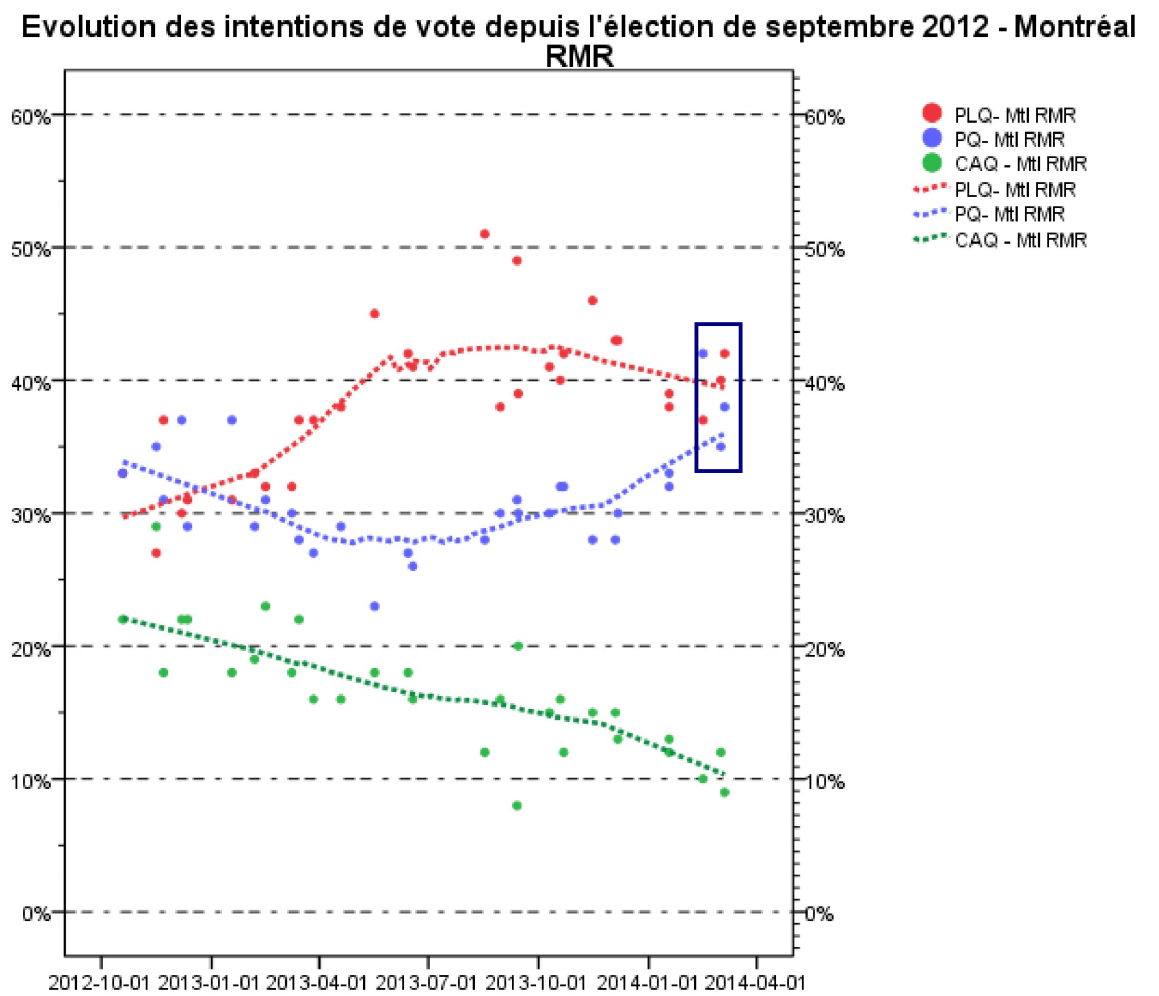
L'évolution des intentions de vote au Québec

Telles que publiées (n ≈ 1000)



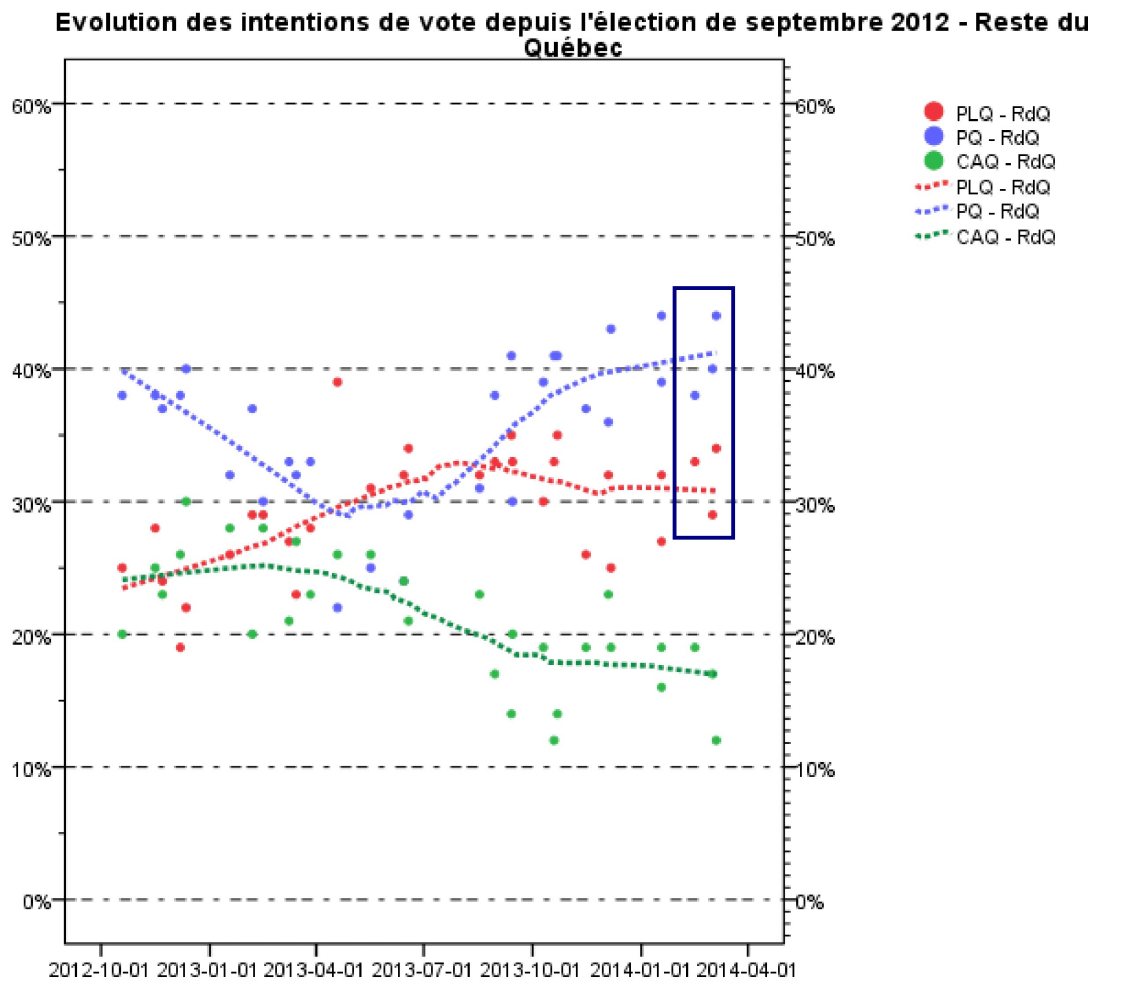
L’évolution des intentions de vote dans la RMR de Montréal

Effet de la taille. N moyen= 436



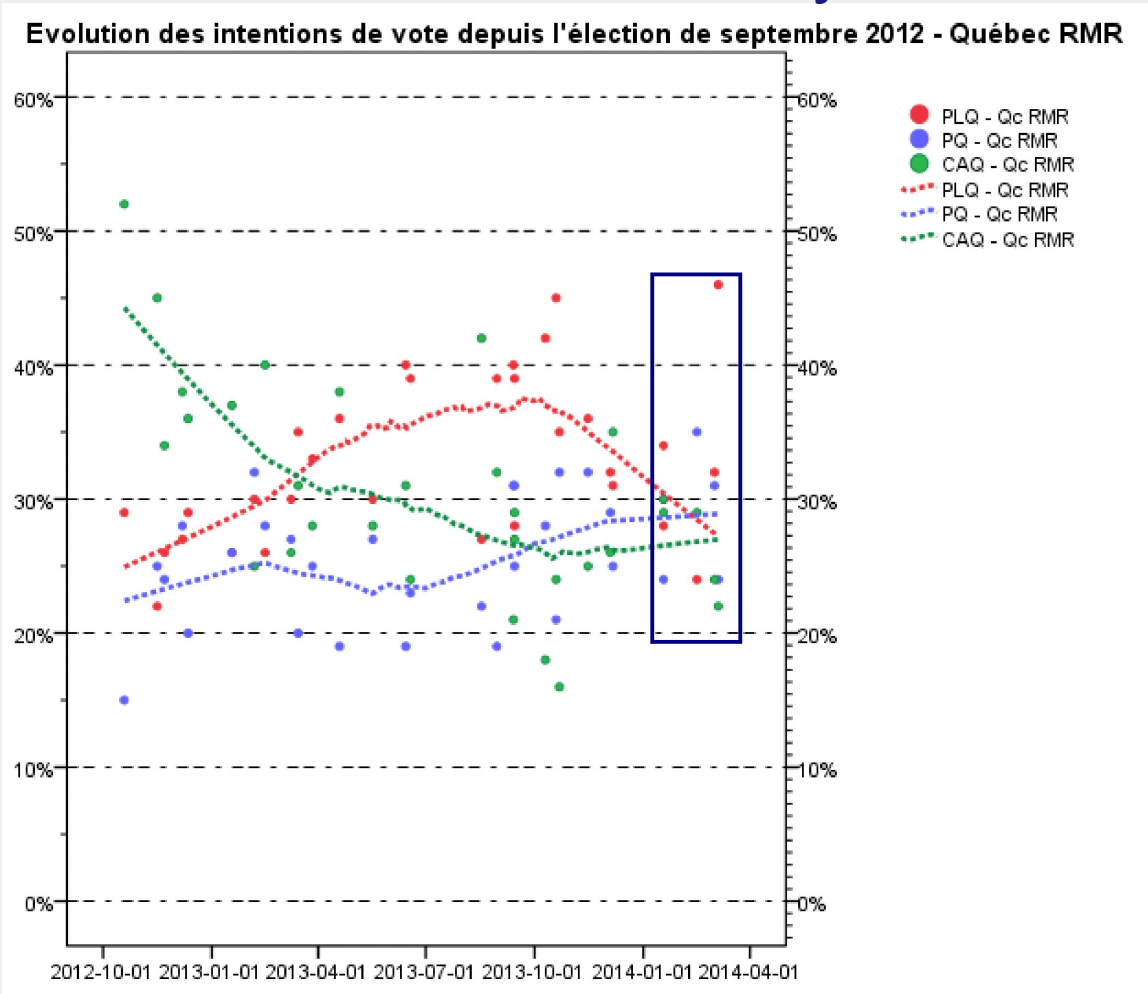
L’évolution des intentions de vote dans le Reste du Québec

Effet de la taille. N moyen= 282



L’évolution des intentions de vote dans la RMR de Québec

Effet de la taille. N moyen= 196



Etes-vous capable de calculer la marge d'erreur et l'intervalle de confiance de la proportion de jaunes dans un sac de M&Ms?

- Disons 15% de jaunes
- Dans un sac de 250 M&M?

La marge d’erreur maximale

Et un peu plus...

- Pourquoi lit-on dans la présentation des sondages électoraux dans les journaux...
  - “La marge d’erreur est de 3,1%, 19 fois sur 20" alors que plusieurs pourcentages différents sont présentés (des pourcentages d’intention de vote pour différents partis, par exemple).
- Réponse: Parce qu’il s’agit de la marge d’erreur ***maximale***, celle que l’on obtient pour une ***proportion de 50%***.

La marge d'erreur maximale

Si on calcule la marge d'erreur pour une proportion p=50% pour un échantillon de 1000 personnes...

- On obtient:

$$e = 1{,}96 * \sqrt{\frac{0{,}5*(1-0{,}5)}{1000}} = 1{,}96 * \sqrt{\frac{0{,}25}{1000}} = 1{,}96 * \sqrt{0{,}00025} = 1{,}96 * 0{,}0158$$

$$e = 0{,}031 \quad et \quad e\% = e*100 = 3{,}1\%$$

ET donc, la marge d'erreur ***maximale*** est de 3,1% pour 1000 répondants dans l'échantillon. Et la proportion se situera 19 fois sur 20 entre 50% - 3,1% et 50% + 3,1% [46,9; 53,1].

La marge d'erreur maximale... simplifiée

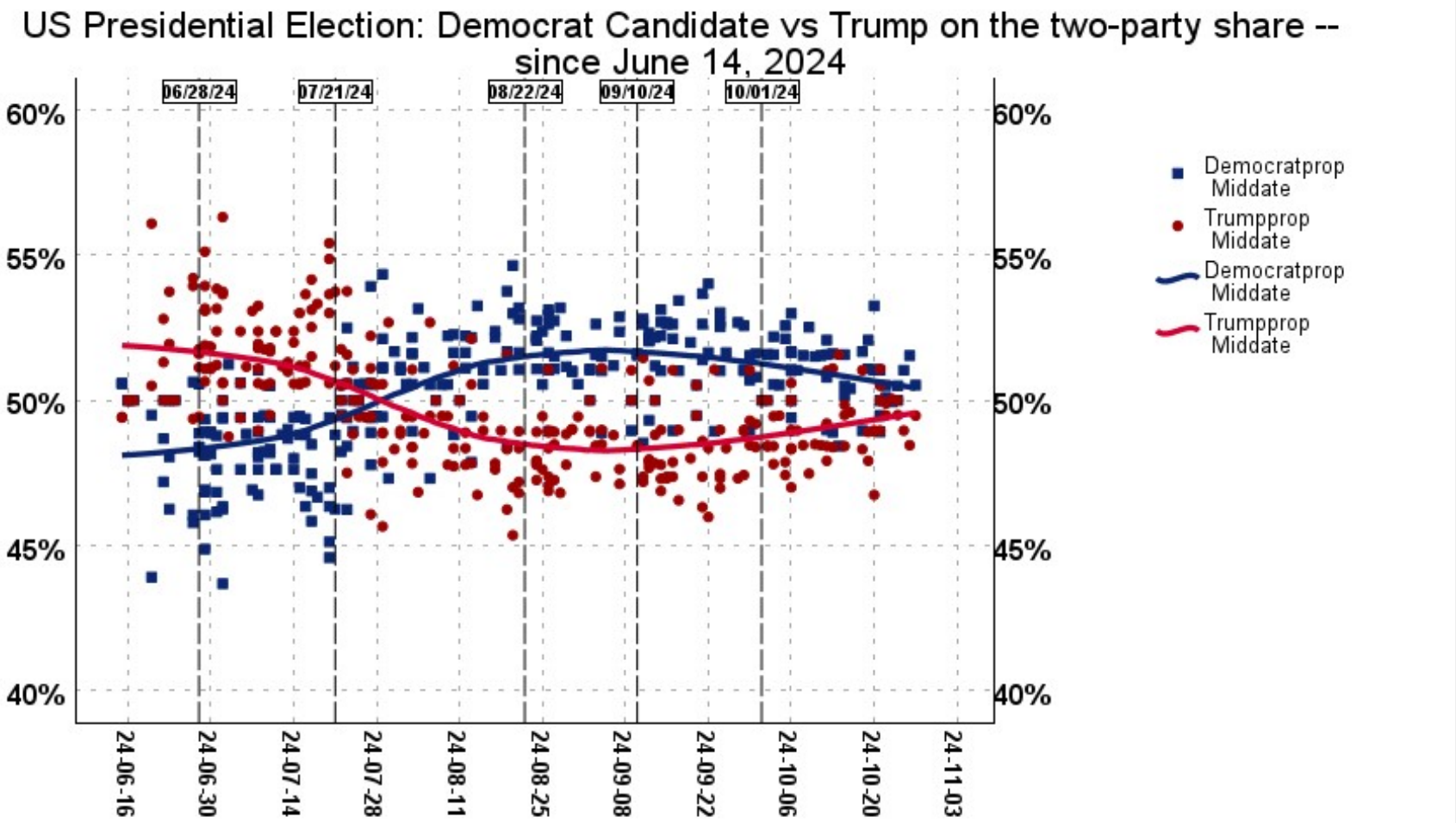
Si on calcule la marge d'erreur à 50% pour un échantillon de n personnes...

- On peut simplifier la formule de la façon suivante:

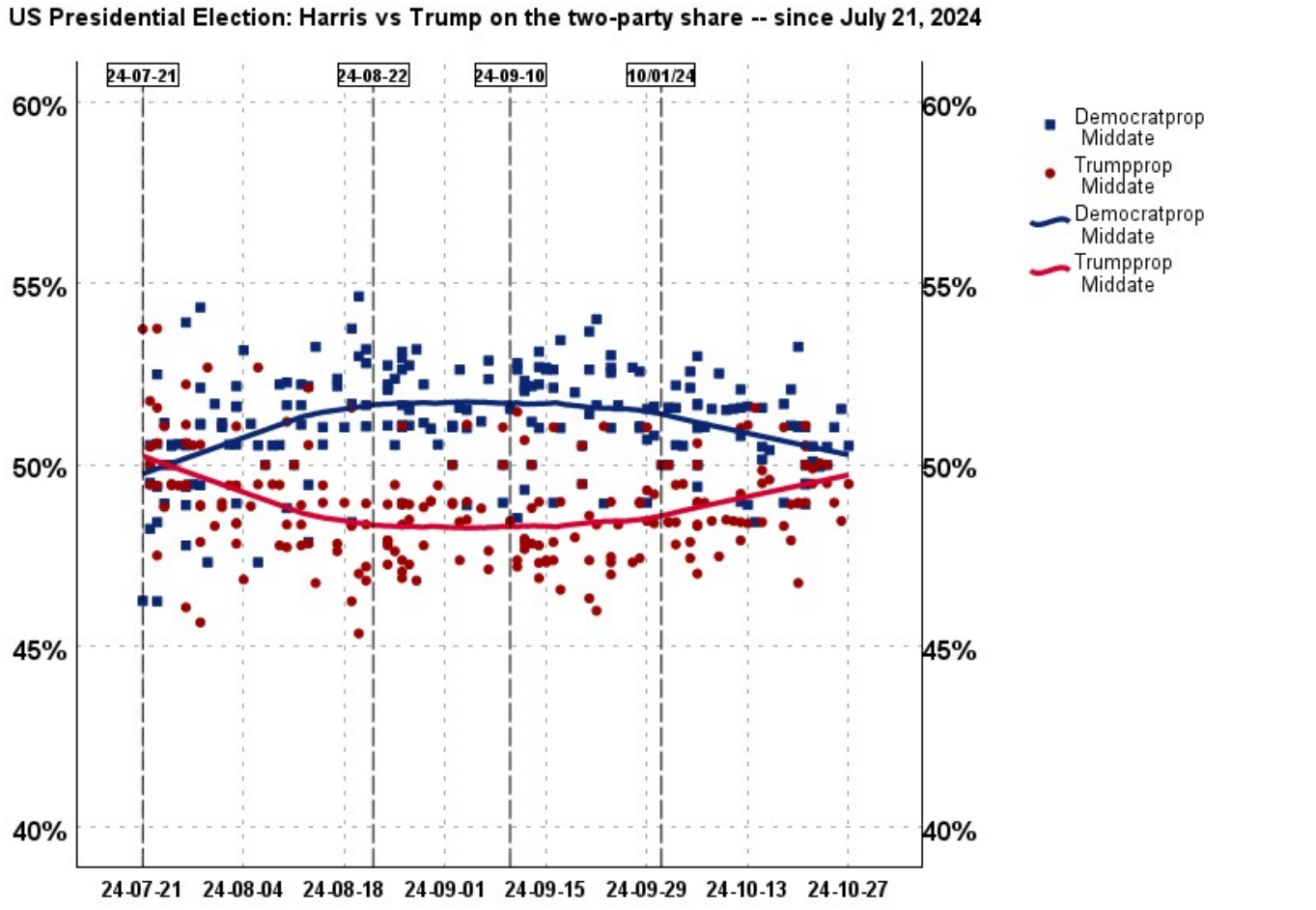
$$e = 1{,}96 * \frac{\sqrt{0{,}25}}{\sqrt{n}} = (1{,}96 * 0{,}5) * \frac{1}{\sqrt{n}} \approx \frac{0{,}98}{\sqrt{n}} \approx \frac{1}{\sqrt{n}}$$

On utilise la marge d'erreur maximale quand on parle de la marge d'erreur d'un sondage, d'un échantillon, et non pas de celle d'un résultat spécifique.

Evolution intention de vote: Élection présidentielle américaine



Evolution intention de vote: Élection présidentielle américaine



La marge d’erreur pour une population finie

La marge d'erreur pour une population finie

Qu'arrive-t-il si la population est petite? Est-ce que la marge d'erreur est la même?

- On parle de marge d'erreur pour une population finie lorsque la ***population*** est moins grande que 20 fois l'échantillon.
  - Exemple: échantillon de 1,000 répondants sur une population de 20,000 (une personne sur 20), ou "pire", de 1000 sur 3000, soit une personne sur 3.
- La formule de la marge d'erreur pour population finie permet de tenir compte du fait que la **population** est plus petite: la marge d'erreur sera donc réduite.

La marge d'erreur pour une population finie (+ petite que 20 fois l'échantillon)

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

- Et donc, pour une proportion de 0,5 (50%) et une taille d'échantillon (n) de 1000 personnes mais une ***population (N) de 5000*** personnes:

$$e = 1{,}96 * \sqrt{\frac{0{,}5*(1-0{,}5)}{1000}} * \sqrt{\frac{5000-1000}{5000-1}} = 0{,}031 * \sqrt{\frac{4{,}000}{4{,}999}} = 0{,}031 * \sqrt{0{,}8} = 0{,}031 * 0{,}89 = 0{,}0277$$

La marge d'erreur est moins grande pour une population finie - à taille d'échantillon égale, soit 2,77%, que pour une population infinie, soit 3,1%

La marge d’erreur de la différence entre deux proportions

Comment savoir si deux proportions sont significativement différentes l'une de l'autre?

- Une façon simple consiste à vérifier si les intervalles de confiance de deux proportions se chevauchent, par exemple...
  - Si on compare une proportion de 45% sur un échantillon de 800 (e=3,4%, comme calculé précédemment) et une proportion de 35% sur le même échantillon (e=3,3%), on obtient les intervalles respectifs suivants:
  - De 41,6% à 48,4% et de 31,7% à 38,3%.
  - Comme le maximum de l'intervalle de 35% (38,3%) est plus petit que le minimum de l'intervalle de 45% (41,6%), on conclut que les deux proportions sont statistiquement différentes (**les intervalles ne se recoupent pas)**. C'est la manière simple et conservatrice de calculer. C'est l'équivalent de multiplier la marge d'erreur par 2.

Marge d'erreur pour une différence de proportion dans un même échantillon

- Toutefois, pour la différence entre deux pourcentages pris sur le même échantillon, la marge d'erreur de la différence, moins conservatrice, devrait se calculer plutôt ainsi:

$$e_{diff} = Z_{\alpha} * \sqrt{\frac{(p_1+p_2)-(p_1-p_2)^2}{n}} = 1{,}96 * \sqrt{\frac{(,45+,35)-(,45-,35)^2}{800}}$$

Ce qui donne 0,061 ou 6,1%

- Où p_1 est la première proportion, p_2 est la deuxième et n est la taille de l'échantillon.
- ***Il faut donc une différence de 6,1% entre les deux proportions*** pour que l'on considère la différence significative. Il s'agit d'un test moins conservateur que celui présenté à la page précédente (6,8%).
  - Comme 45%-35%=10% est plus grand que 6,1%, la différence est significative.

Marge d'erreur pour une différence de proportions entre 2 échantillons différents

- Pour la différence entre deux proportions prises sur des échantillons différents (comme la différence entre les résultats de deux sondages), la marge d'erreur se calcule ainsi:

$$e_{diff} = Z_{\alpha} * \sqrt{\frac{p1*(1-p1)}{n1} + \frac{p2*(1-p2)}{n2}}$$

Ce qui s'approche de:

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

- Où "p" est égal à $(p_1+p_2)/2$, soit la moyenne des deux proportions dont on veut estimer la différence et n est la moyenne de n1 et n2
- Cette erreur s'approche de 1,4*e.

Avertissement

- Notons que les formules vues dans ce cours ne sont valides que pour les échantillons aléatoires simples avec probabilités égales de sélection,...
- Dans le cadre du cours, nous les utilisons comme formules générales “de base” puisque nous ne pouvons pas regarder toutes les formules spécifiques qui s’appliquent dans les diverses situations (cours de statistiques).
- Au besoin, il faudrait trouver ces formules et les appliquer. En pratique, certains logiciels font le calcul automatiquement (Stata, SPSS, R, etc).

Ressources

- Sur le site de Circum (https://ssl.circum.com/index.cgi?fr:news), vous trouverez un calculateur de marge d'erreur "**Echanticalc**" et un calculateur *d'effet de plan*, qui permet de tenir compte du fait que les probabilités de sélection sont inégales.
- Sur le site de "Si la tendance se maintient", vous trouverez un calculateur pour la différence entre deux proportions: http://www.tooclosetocall.ca/p/is-it-significant-est-ce-significatif.html.

Résumé des “notions”

- ℓ : marge d’erreur
- ℓ_{max} : marge d’erreur maximale quand p=0,5 (50%)
- ℓ_{diff} : marge d’erreur d’une différence entre deux proportions
  - D’un même échantillon.
  - De deux échantillons différents.
- Z_α: Valeur de Z pour un seuil de confiance 1-α
- p: proportion de présence d’une caractéristique dans l’échantillon
- n: taille de l’échantillon final
- N: taille de la base échantillonnale