Chapitre 6:

Les proportions

6.1. Ecart type pour les proportions

Considérons le cas d'un sondage politique.

Ici, on ne désire plus estimer la moyenne d'une grandeur sur une population, mais la proportion des individus de cette population qui se rangent dans une catégorie (p.ex., qui déclarent voter pour le P.U.B.).

Exemple

Un institut de sondage interroge un échantillon représentatif de 200 électeurs, qui ont le choix entre 4 partis politiques.

Les résultats du sondage sont les suivants:

parti	nombre d'intentions de vote
PUB	24
PET	35
PAF	69
PIF	61
ne se prononcent pas	11

Notons p la proportion d'individus d'une classe dans l'échantillon

Si n_1 est le nombre d'individus dans la classe 1 et n le nombre total d'individus dans l'échantillon, on a :

$$p_1 = \frac{n_1}{n}$$

et de même

$$p_2 = \frac{n_2}{n}, \cdots$$

Dans notre sondage, les proportions sont les suivantes:

parti	proportion
PUB	0,120
PET	0,175
PAF	0,345
PIF	0,305
n.s.p.	0,055

Si l'échantillon est représentatif, la proportion *p* dans l'échantillon est une approximation de la proportion π dans la population.

Pour des échantillons suffisamment grands, les proportions suivent une loi normale, avec un écart type d'échantillon de:

$$\sigma_p = \sqrt{\frac{\pi\,(1 - \pi)}{n}}$$

En général, la proportion π dans la population n'est pas connue. On la remplace alors par la proportion *p* dans l'échantillon

$$\sigma_p \cong \sqrt{\frac{p(1-p)}{n}}$$

Les proportions obéissent à des lois comparables à celles des moyennes.

Une différence importante est que l'écart type peut être calculé à partir des proportions (pour les moyennes, il devait être connu par ailleurs).

Nous pouvons donc calculer les intervalles de confiance à 95 % sur les intentions de vote.

$$\sigma_p(\text{PUB}) = \sqrt{\frac{0{,}12\,(1 - 0{,}12)}{200}} = 0{,}023$$

$$\sigma_p(\text{PET}) = \sqrt{\frac{0{,}175\,(1 - 0{,}175)}{200}} = 0{,}027$$

$$\sigma_p(\text{PAF}) = \sqrt{\frac{0{,}345\,(1 - 0{,}345)}{200}} = 0{,}034$$

$$\sigma_p(\text{PIF}) = \sqrt{\frac{0{,}305\,(1 - 0{,}305)}{200}} = 0{,}033$$

Les intervalles de confiance à 95 % sont de $2\sigma_P$

Les résultats du sondage sont les suivants:

parti	intentions de vote
PUB	12,0 ± 4,6 %
PET	17,5 ± 5,4 %
PAF	34,5 ± 6,8 %
PIF	30,5 ± 6,6 %

6.2. Les proportions sont des moyennes

Considérons une élection opposant deux partis A et B.

Considérons la grandeur *x* = nombre de voix qu'un électeur apporte au parti B.

C'est une variable discrète qui peut prendre deux valeurs :

$x = 0$	si l'électeur vote pour A
$x = 1$	si l'électeur vote pour B

Soient

n_A	le nombre d'électeurs votant pour A
n_B	le nombre d'électeurs votant pour B
n	le nombre total d'électeurs
p	la proportion d'électeurs votant pour B

Calculons la valeur moyenne de x :

$$\bar{X} = \frac{1}{n}\sum x = \frac{1}{n}\,(\underbrace{0 + 0 + \ldots}_{n_A} + \underbrace{1 + 1 + \ldots}_{n_B})$$

$$\bar{X} = \frac{n_B}{n} = p$$

La proportion est donc la moyenne de *x*.

Calculons l'écart type sur x (ou plutôt son carré, appelé variance) :

$$\sigma^2 = \frac{1}{n} \sum (x - \overline{X})^2$$

$$\sigma^2 = \frac{1}{n} [\underbrace{(0 - p)^2 + (0 - p)^2 + \ldots}_{n_A} + \underbrace{(1 - p)^2 + (1 - p)^2 + \ldots}_{n_B}]$$

$$\sigma^2 = \frac{1}{n} [\, n_A p^2 + n_B (1 - p)^2 \,]$$

$$\sigma^2 = \frac{n_A}{n} p^2 + \frac{n_B}{n} (1 - p)^2$$

$$\sigma^2 = (1 - p)\, p^2 + p\, (1 - p)^2$$

$$\sigma^2 = p\,(1 - p)\,[\, p + 1 - p \,]$$

On a donc

$$\sigma = \sqrt{p\,(1 - p)}$$

Et donc,

$$\sigma_p = \sigma(p) = \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

$$\Longrightarrow \sigma = \sqrt{\frac{p\,(1 - p)}{n}}$$

6.3. Exemples

Exemple 1 :

$n = 100 \quad n_1 = 60 \quad n_2 = 40$

$$p_1 = \frac{n_1}{n} = \frac{60}{100} = 0{,}6 \qquad p_2 = \frac{n_2}{n} = \frac{40}{100} = 0{,}4$$

$$\left.\begin{aligned} \sigma_1 &= \sqrt{\frac{p_1\,(1 - p_1)}{n}} = \sqrt{\frac{0{,}6 \times 0{,}4}{100}} = 0{,}049 \\ \sigma_2 &= \sqrt{\frac{p_2\,(1 - p_2)}{n}} = \sqrt{\frac{0{,}4 \times 0{,}6}{100}} = 0{,}049 \end{aligned}\right\} \text{ Pourquoi = ? } *$$

* Quand il n'y a que deux choix possibles, et pas d'abstentions, on a $n_2 = n - n_1$ et l'incertitude sur n_2 est forcément la même que sur n_1. Ce n'est plus vrai à partir de 3 choix.

On ne peut pas calculer

$$\sigma_D = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2}$$

car les 2 échantillons ne sont pas indépendants !

Le parti 1 gagne les élections si $p_1 > 0,5$

$\delta = 0,6 - 0,5 = 0,1$

$\sigma = 0,049$

$$z_0 = \frac{\delta}{\sigma} = \frac{0,1}{0,049} = 2.04$$

table $\rightarrow$ Prob = 0,021

$\rightarrow$il y a 2,1 % de chances que $p_1 < 0,5$

$\rightarrow$il y a 100-2,1 = 97,9 % de chances que le parti 1 remporte les élections

Exemple 2 :

On constate un défaut dans 20 % des voitures d'un modèle. Un garagiste, qui a vendu 50 voitures de ce modèle, fait revenir tous ses clients afin de remplacer une pièce aux voitures défectueuses. Pour cela, il a commandé 12 pièces de rechange.

Quelle est la probabilité qu'il n'ait pas suffisamment de pièces ?

Solution :

La proportion de voitures défectueuses vaut $\pi = 0,2$.

Dans l'échantillon de 50 voitures, on s'attend à la même proportion, avec un écart type :

$$\sigma_p = \sqrt{\frac{\pi\,(1 - \pi)}{n}} = \sqrt{\frac{0,2 \times 0,8}{50}} = 0,057$$

Il n'aura pas suffisamment de pièces si le nombre de voitures défectueuses est supérieur à 12, ou au moins égal à 13.

On a le choix entre deux critères :

$$p > \frac{12}{50} \qquad \text{ou} \qquad p \geq \frac{13}{50}$$

Dans ce cas, on obtient un meilleur résultat en appliquant la *correction de continuité* et en choisissant 12,5 plutôt que 12 ou 13.

Nous retiendrons donc

$$p > \frac{12,5}{50} = 0,25$$

Pour qu'il n'ait pas assez de pièces de rechange, il faut donc que la proportion s'écarte de la valeur moyenne de plus de

$$\delta = 0,25 - 0,2 = 0,05$$

On obtient donc :

$$z_0 = \frac{\delta}{\sigma} = \frac{0,05}{0,057} = 0,88$$

La probabilité d'avoir un écart au moins aussi élevé est, d'après la table de la loi normale, de 0,189.

Il y a donc 18,9 % de risques que le garagiste n'ait pas assez de pièces de rechange.

6.4. Illustration : test de la fiabilité des horoscopes

Lors de deux leçons, nous organisons un test destiné à mesurer la fiabilité des horoscopes.

Dans ce but, nous distribuons des feuilles reprenant les horoscopes de la semaine précédente.

Chaque étudiant présent lit ces horoscopes et indique :

- son signe astrologique ;
- lequel de ces horoscopes correspond le mieux à ce qui lui est arrivé lors de la semaine écoulée.

Pour le premier test, les signes astrologiques sont indiqués.

Dans le second test, ces signes ne sont pas indiqués et l'ordre en est modifié.

Ce test a pour but de répondre à deux questions :

- les horoscopes sont-ils fiables ?
- la connaissance du signe a-t-elle une influence sur les réponses des individus testés ?

Nous désignerons par « coïncidences positives » (CP) les cas où l'étudiant a reconnu son signe, c'est-à-dire les cas où l'horoscope qu'il a sélectionné correspond bien à son signe.

Les résultats des tests effectués de 1998 à 2008 sont résumés dans le tableau suivants :

année	signes connus		signes inconnus	
	n	CP	n	CP
1998-2001	145	46	275	22
2002	71	9	71	9
2003	57	10	72	5
2004	81	15	44	2
2007	61	12	58	6
2008	55	6	64	5
2004	81	15	44	2
total	**470**	**79**	**584**	**49**

6.4.1. Test de la fiabilité des horoscopes.

Nous considérons l'ensemble des quatre années et retenons les tests où les signes n'étaient pas connus, afin d'éviter des biais éventuels.

Nous avons un échantillon de 584 réponses, avec 49 coïncidences positives.

Si ces coïncidences positives étaient dues au hasard uniquement, c'est-à-dire si chaque individu répondait au hasard, il aurait une chance sur 12 de choisir l'horoscope correspondant à son signe.

Par le hasard seul, nous nous attendrions donc à

$$\frac{584}{12} = 48{,}7 \quad \text{CP}$$

Or, nous avons 49 CP, ce qui est pratiquement égal au nombre attendu par l'action du hasard.

C'est très mauvais signe pour la fiabilité des horoscopes !

Nous allons cependant utiliser nos connaissances en statistique pour analyser ces tests de manière plus quantitative.

(a) le nombre de coïncidences positives est compatible avec l'action du hasard seul.

Dans le cas d'une répartition au hasard, le nombre de CP doit être, en moyenne, $\frac{1}{12}$ du nombre de réponses.

La proportion de CP, p_+, vaut donc :

$$p_+ = \frac{1}{12} = 0{,}083$$

avec un écart type

$$\sigma_+ = \sqrt{\frac{p_+ (1 - p_+)}{n}} = \sqrt{\frac{0{,}083 \times 0{,}917}{584}} = 0{,}011$$

L'intervalle de confiance à 95 % vaut donc :

$$0{,}083 \pm 0{,}022$$

ou encore :

$$[0{,}061 ; 0{,}105]$$

La valeur obtenue par l'étude de notre échantillon vaut : $p_+ = \frac{49}{584} = 0{,}084$

Elle se trouve dans l'intervalle de confiance pour une répartition due au hasard.

On peut donc conclure que la petite différence entre la valeur mesurée et la valeur attendue est parfaitement compatible avec le hasard : c'est ce qu'on appelle une *fluctuation statistique.*

Exemple de fluctuation statistique.

Si on lance une pièce de monnaie, on s'attend à avoir, en moyenne, autant de « pile » que de « face ».

Sur 100 lancers, on n'aura que rarement 50 « pile » et 50 « face » exactement. Les écarts par rapport à ce nombre moyen sont les fluctuations statistiques.
ex : 47 pile et 53 face,
52 pile et 48 face,…

(b) Avec quelle confiance pouvons-nous conclure à la non-fiabilité des horoscopes ?

Nous devons tout d'abord définir ce que nous entendons par fiabilité des horoscopes.

Si l'astrologie était une science exacte, elle devrait être capable de prédire avec certitude ce qui va nous arriver.

Toutefois, nous ne lui en demanderons pas tant.

Nous dirons que les horoscopes sont fiables à 50 % si les prédictions concernant notre signe sont celles qui correspondent le mieux à ce qui nous arrive, dans au moins un cas sur deux.

Dans ce cas, au moins la moitié des individus devraient reconnaître leur signe.

Remarque : cette définition est très peu contraignante pour l'astrologie. En effet :

- nous ne lui demandons pas de prédire avec précision ce qui va nous arriver, mais seulement que la prédiction concernant notre signe soit la plus proche de ce qui va nous arriver, parmi les 12 prédictions.
- nous ne demandons pas que cela se produise pour tous les individus, mais seulement pour la moitié d'entre eux.

Soit p_+ la proportion des individus qui reconnaissent leur signe.

Dans notre échantillon, nous avons :

$$p_+ = \frac{49}{584} = 0{,}084$$

avec un écart type :

$$\sigma_+ = \sqrt{\frac{p_+ (1 - p_+)}{n}} = \sqrt{\frac{0{,}084 \times 0{,}916}{584}} = 0{,}011$$

Pour que les horoscopes soient fiables à 50 %, il faudrait, dans la population, une proportion $p_+ \geq 0.5$, donc un écart minimum avec notre valeur d'échantillon :

$$\delta_+ = 0{,}5 - 0{,}084 = 0{,}416$$

et donc :

$$Z_0 = \frac{\delta_+}{\sigma_+} = \frac{0{,}416}{0{,}011} \cong 38\ !$$

Cette valeur est si grande qu'elle ne figure pas dans notre table de la loi normale.

En fait, il n'y a pas une chance sur des milliards de milliards pour que les horoscopes testés soient fiables à 50 %.

Notre échantillon nous permet d'exclure cette hypothèse avec une certitude quasi absolue.

Les horoscopes pourraient-ils être fiables à 25 % ?

Pourrait-il y avoir une chance sur 4 pour que la prédiction qui correspond le mieux à un individu soit celle de son signe ?

Dans ce cas, au moins un quart des individus devraient reconnaître leur signe.

Il faudrait donc $p_+ \geq 0.25$, et donc un écart

$$\delta_+ = 0{,}25 - 0{,}084 = 0{,}166$$

et :

$$Z_0 = \frac{\delta_+}{\sigma_+} = \frac{0{,}166}{0{,}011} \cong 15{,}1$$

Cette valeur est, une fois de plus, en dehors de la table de la loi normale.

Notre test nous permet d'exclure avec une quasi certitude que les horoscopes testés soient fiables une fois sur 4.

Tester une fiabilité plus faible n'a pas beaucoup de sens car :

- être fiable moins d'une fois sur 4, c'est plutôt être non fiable.
- le hasard seul donne une fiabilité d'une fois sur 12$\rightarrow$on risque évidemment de trouver que les horoscopes sont fiables une fois sur 12 !

En résumé :

Nous pouvons conclure que la fiabilité des horoscopes testés est nulle, puisque l'on obtiendrait le même résultat en choisissant les signes au hasard.

Si vous lisez les horoscopes, rien ne sert de connaître votre signe. Les prévisions des autres signes s'appliquent tout aussi bien (ou plutôt : tout aussi mal) à vous !

6.4.2. La connaissance du signe a-t-elle influencé les réponses ?

Nous allons tâcher de déterminer si les individus testés se sont laissé influencer par la connaissance de leur signe, lorsque celui-ci était indiqué.

Dans les tests avec signes connus, nous avons 79 coïncidences positives sur 470 réponses, soit une proportion:

$$p_c = \frac{79}{470} = 0{,}168$$

avec un écart type:

$$\sigma_c = \sqrt{\frac{0{,}168 \times 0{,}832}{470}} = 0{,}017$$

Avec les signes inconnus, nous avons 49 coïncidences positives sur 584 réponses, soit une proportion:

$$p_i = \frac{49}{584} = 0,084$$

avec un écart type:

$$\sigma_i = \sqrt{\frac{0,084 \times 0,916}{584}} = 0,011$$

On a donc une proportion plus grande de coïncidences positives lorsque les signes sont connus, ce qui laisse supposer que certains individus se sont laissé influencer par la connaissance de leur signe .

Cette différence est-elle statistiquement significatives ?

Avec quelle confiance pouvons-nous affirmer que cette différence ne peut pas être due à l'action du hasard (fluctuation statistique).

Nous pouvons supposer que les deux échantillons sont indépendants car nous ne voyons pas comment la réponse à un des test pourrait influencer la réponse à l'autre.

Nous avons une différence de proportion:

$$\delta = p_c - p_i = 0,168 - 0,084 = 0,084$$

entre les CP avec signes connus et inconnus.

L'écart type sur cette différence vaut:

$$\sigma_\delta = \sqrt{\sigma_c^{\,2} + \sigma_i^{\,2}} = \sqrt{0,017^2 + 0,011^2} = 0,020$$

Nous obtenons donc:

$$Z_0 = \frac{\delta}{\sigma_\delta} = \frac{0,084}{0,020} \cong 4,0$$

D'après la loi normale, la probabilité qu'un tel écart soit dû au hasard est de :

$$0,00003 = 0,003\ \%$$

Nous pouvons donc conclure avec 99,997 % de confiance que la connaissance du signe a effectivement influencé les réponses.

Ce résultat illustre l'importance de réaliser les tests "à l'aveugle", sans que les sujets testés puissent se laisser influencer par la connaissance d'informations de nature à influencer le résultat. Même en essayant de ne pas tenir compte de ces informations, on risque fort de se laisser influencer.