Université Abou Bakr Belkaid

Département de Biologie

L3 AACQ S2 (2019-2020)

Chapitre I : Correlation et regression

Chapitre II : Estimation

[1]

1. D^r *A.BOUKENKOUL*

Table des matières

Chapitre 1

Correlation et régression

Remarque 1.0.1. *Ce chapitre a été vu en classe*

Chapitre 2

Estimation des paramètres inconnus

On veut étudier un caractère A dans une population Ω à laquelle on n'a pas accès car elle comporte un ou des paramètres inconnus. $\grave{A}$ partir d'un échantillon on n'a pas des certitudes mais des estimations.
Conditions d'approximation de la loi Binomiale par la loi normale :

Les conditions nécessaires pour approximer une loi binomiale par une loi normale sont :

1. l'échantillon doit $\hat{e}$tre aléatoire simple.
2. les conditions pour une loi binomiale sont satisfaites.Á savoir, il y a un nombre fini n de répititions, les essais sont indépendants, il y a deux catégories de résultats et les probabilités sont constantes pour tous les essais.
3. la loi normale peut être utilisée pour approximer la loi des proportions d'échantillon si les conditions $np \geq 5$ et $nq \geq 5$ sont toutes les deux satisfaites.

2.1 Estimer une proportion dans une population

Objectifs :

1. Comprendre ce que sont les intervalles de confiance, leur signification et leur utilité.

2. Cnstruire un intervalle de confiance pour l'estimation d'une proportion.
3. Interprétation corrcte d'un intervalle de confiance.

2.1.1 Notations pour les proportions

p : proportion de succès dans toute la population.
$\hat{p} = \frac{x}{n}$: proportion de x succès dans un échantillon de taille n.
$\hat{q} = 1 - \hat{p}$: proportion de $(n - x)$ échecs dans un échantillon de taille n.

2.1.2 Quelques définitions

1. Une estimation ponctuelle est une valeur unique utilisée pour approximer le paramètre d'une population.
2. Un intervalle de confiance, parfois noté IC, est un intervalle de valeurs utilisé pour estimer la vraie valeur d'un paramètre d'une population.
3. Le niveau de confiance parfois noté NC est la probabilité qui est la proportion d'un grand nombre de fois où l'IC contient le paramètre de la population si on répète l'estimation un grand nombre de fois.

2.1.3 Valeurs critiques

Une valeur critique est un grand nombre sur la frontière séparant les statistiques d'échantillon qui peuvent vraisemblablement survenir de celles qui ne le peuvent pas. Le grand nombre $Z_{\frac{\alpha}{2}}$ est une valeur critique qui est des aires de $\frac{\alpha}{2}$ pour la loi normale standard.

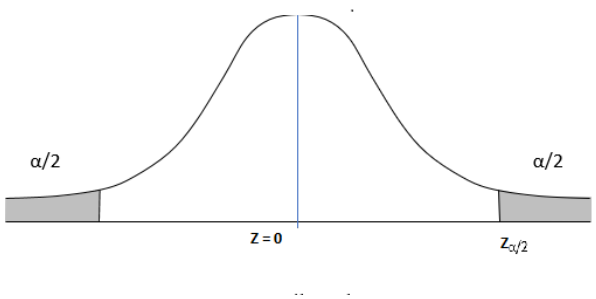

FIGURE 2.1 – Lue dans la table de la loi normale(correspond à une aire de $1 - \frac{\alpha}{2}$)

Exemple 2.1.1. *Trouvez la valeur critique $Z_{\frac{\alpha}{2}}$ qui correspond à l'intervalle de confiance à 95%*

Solution 1. *Attention : il ne faut pas chercher 0,95 dans le table de la loi normale.*
Un IC à 95% correspend à
$\alpha = 5\% = 0,05$ *soit* $\frac{\alpha}{2} = 0,025$ *donc c'est la valeur* $1 - 0,025 = 0,975$ *qu'il faut chercher. On trouve alors* $Z_{\frac{\alpha}{2}} = 1,96$

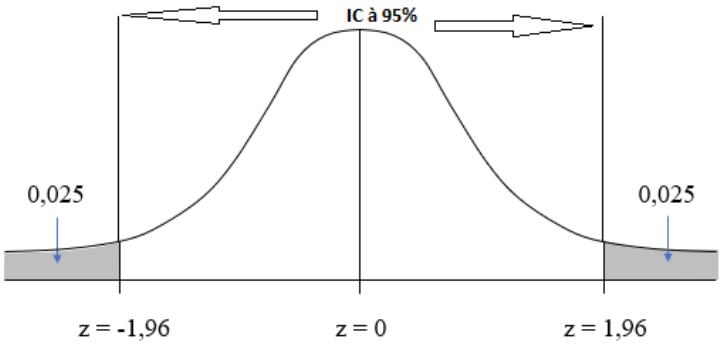

FIGURE 2.2 – L'aire totale à la gauche de la frontière 1.96 est de 0.975

valeurs usuelles :

Niveau de confiance (1-α)	risque(α)	valeurs critiques $Z_{\frac{\alpha}{2}}$
90%	10% = 0,10	1,645
95%	0,05	1,96
99%	0,01	2,575

Marge d'erreur : appelée aussi l'erreur maximale de l'estimation, on la note par E et on a :

$$E = Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

L'intervalle de confiance (IC) pour la proportion p de Ω :

$$\hat{p} - E < p < \hat{p} + E \text{ où } E = Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$[\hat{p} - E;\ \hat{p} + E]$$

où $\hat{p}\bar{+}E$.

Exemple 2.1.2. *Dans l'une des experiences génétiques de Mendel, on a noté que 580 pois ont été obtenus et que 26,2% avaient des gousses jaunes. On note que l'estimation ponctuelle de la proportion p de la population est 0,262. Utilisez ces résultats pour répondre aux questions suivantes :*

1. *Trouvez la marge d'erreur qui correspond à un intervalle de confiance à 95%.*
2. *Trouvez l'intervalle de confiance à 95% de la proportion p de la population.*
3. *À partir de ces résultats, que peut-on conclure sur la théorie de Mendel qui déclare que le pourcentage de pois à grousses jaunes devrait être égal à 25% ?*

Solution 2. *Nous devons d'abord vérifier que les conditions requises sont satisfaites*

1. *L'echantillon est aléatoire simple (vue la conception de l'expérience de Mendel)*
2. *Les conditions d'une experience binomaile sont satisfaites parceque :*

   (a) *Il y a un nombre fixe déssais (580).*

   (b) *les essais sont indépendants (la couleur d'une gousse ne modifie pas la couleur d'une autre gousse).*

   (c) *Il y a 2 catégories (jaune, pas jaune) et la probabilité d' être jaune est constante .*

3. *Finalement, utilisons* $n = 580$ *et* $\hat{p} = 0,262$, $\hat{q} = 1 - \hat{p}$. *Pour montrer que :*

$$n\hat{p} \geq 5 \text{ et } n\hat{q} \geq 5,$$

   Ceci est vérifié car :

$$n\hat{p} = 152 \text{ et } n\hat{q} = 428,$$

   donc la loi normale peut être utilisée pour approximer la loi binomiale.

1. *La marge d'erreur est :*

$$IC : E = Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$
$$avec\ n = 580; \hat{p} = 0,262, \hat{q} = 1 - \hat{p} = 0,738, Z_{\frac{\alpha}{2}} = 1,96$$
$$On\ a : E = 1,96\sqrt{\frac{0,262\times 0,738}{580}} = 0,035787$$

2. *L'intervalle de confiance :*

$$\hat{p} - E < p < \hat{p} + E$$
$$0,262 - 0,035787 < P < 0,262 + 0,035787$$
$$0,226 < P < 0,298 \textit{(arrondi á 3 décimals}$$

on peut aussi exprimer l'intervalle de confiance sous la forme $0,262 \mp 0035787$

3. *À partir de ces résultats nous sommes sûrs à 95% que les limites 22,6% et 29,8% contiennent le vrai pourcentage de pois à gousses jaunes. Comme cet intervalle de confiance contient la valeur 25%, la valeur de Mendel ne peut pas être considérée comme fausse.*
   Les résultats otenus ne contiennent donc aucune preuve significative contre l'hypothèse de 25% de Mendel.

Remarque :

1. *Interprétation correcte : Nous avions confiance à 95% que l'intervalle de confiance* $[0,226\,;\,0,298]$ *contient la vraie valeur p.*
2. *Interprétation incorrecte : Il y a 95% de chance que l'intervalle de confiance* $[0,226\,;\,0,298]$ *contient la vraie valeur de p.*

Taille d'échantillon pour estimer la proportion p

Définition 2.1.1. *la taille d'échantillon est :*

1. *Quand une estimation $\hat{p}$ est connue au paravant :*
$$n = \frac{z^2_{\frac{\alpha}{2}} . \hat{p}. \hat{q}}{E^2}.$$
2. *Quand aucune estimation $\hat{p}$ est connue au préalable ($\hat{p}$ inconnue) :*
$$n = \frac{z^2_{\frac{\alpha}{2}} . 0,25}{E^2}.$$
Il faut arrondir à l'entier immediatement suprieur

Exemple 2.1.3. *la façon dont nous communiquons est de nos jours très liée à l'utilisation du répondeurs, fax, messageries vocales et E-mail.*
Supposez qu'on veuille déterminer le pourcentage de foyers américains qui utilisent les e-mails. Combien de foyers faut-il enquêter de façon à être sûr à 95% que le pourcentage d'échantillon aura une erreur de moins de 4 points ?

1. *Utilisez ce résultat d'une enquête précédente : en 1997; 16,9% des foyers américains utilisaient les e-mails(d'aprés les données de The world Almanach and Book of Fack).*
2. *Supposez qu'on n'a aucune information préalable suggérant une valeur possible de $\hat{p}$.*

Solution 3.
1. *L'enquête précédente suggère $\hat{p} = 0,169$ donc $\hat{q} = 0,831$ avec un niveau de confiance de 95% nous avons $\alpha = 0,05$ et donc $z^2_{\frac{\alpha}{2}} = 1,96$. La marge d'erreur E vaut $0,04$ (l'équivalent décimal de 4 points) donc :*
$$\begin{cases} n = \dfrac{z^2_{\frac{\alpha}{2}} . \hat{p}. \hat{q}}{E^2}, \\ = \dfrac{(1,96)^2 \times 0,169 \times 0,831}{(0,04)^2} = 337,194 \simeq 338, \end{cases}$$
Il faut enquêter au moins sur 338 foyers séléctionnés aléatoirement
2. *On a toujours $z^2_{\frac{\alpha}{2}} = 1,96$ et $E = 0,04$ mais sans aucune connaisance de $\hat{p}$ et $\hat{q}$ donc :*
$$\begin{cases} n = \dfrac{z^2_{\frac{\alpha}{2}} \times 0,25}{E^2}, \\ = \dfrac{(1,96)^2 \times 0,25}{(0,04)^2} = 600,25 \simeq 601, \end{cases}$$

Interprétation 1. *Pour être sûr à 95% que le pourcentage de notre échantillon est à moins de 4 points de la vraie valeur pour les foyers, nous devons tirer au hasard et interroger 601 foyers.*
Si on compare ce résultat à la valeur 338 trouvée dans la partie a ; on constate que si on n'a pas connaissance d'une étude antérieure, il faut un échantillon plus grand pour obtenir le même résultat qu'avec une valeur estimée $\hat{p}$. Mais utilisons un peu notre bon sens : on sait que l'utilisation des e-mails croît si rapidement que l'estimation de 1997 est trop vieille pour être vraiment utile. Aujourd'hui, il y a substentiellement plus de 16,9% de foyers qui utilisent les e-mails.
De façon réaliste, nous avons besoin d'un échantillon plus grand que 338 foyers. Supposons qu'en ne connaît pas vraiment le taux d'utilisation des e-mails, on doit séléctionner aléatoirement 601 foyers.

L'estimation ponctuelle de p et la marge d'erreur E à partir d'un intervalle de confiance :

Si on connait les limites de l'intervalle de confiance, la proportion de l'échantillon $\hat{p}$ et la marge d'erreur E peuvent être calculées ; comme suit :

- *Estimation ponctuelle de p est :* $\hat{p} = \dfrac{\textit{limite superieure} + \textit{limite inferieure}}{2}$
- *Marge d'erreur :* $E = \dfrac{\textit{limite superieure} - \textit{limite inferieure}}{2}$

Exemple 2.1.4. *L'article high-Dass Nicotine patch theorapy de Dale, Hurt and Al(Journal of the American Medical association, vol.274 ; n $\hat{=}$ 17) inclut cette phrase :"Sur les 71 sujets, 70% se sont abstenus de fumer au bout de 8 semaines (IC à 95% : 58% à 81%)."*
Utilisez cette phrase pour l'estimation ponctuelle $\hat{p}$ et la marge d'erreur E.

Solution 4. *À partir de l'article, l'intervalle de confiance est* $0,58 \leq p \leq 0,81$. *L'estimation ponctuelle $\hat{p}$ est à mi-chemin de ces deux valeurs, soit :*

$$\hat{p} = \frac{\textit{lim Sup} + \textit{lim Inf}}{2} = \frac{0,81 + 0,58}{2} = 0,695$$

La marge d'erreur E est :

$$E = \frac{\textit{lim Sup} - \textit{lim Inf}}{2} = \frac{0,81 - 0,58}{2} = 0,115$$

2.1.4 Exercices

Exercice 2.1.1. *1. Trouvez la valeur critique $Z_{\frac{\alpha}{2}}$ qui correspond aux niveaux de confiance suivants 99% et 98% .*

2. Donnez l'expression de l'intervalle de confiance $0,220 < p < 0,280$ sous la forme $\hat{p} \mp E$.

3. Exprimez l'intervalle de confiance $[0,604; 0,704]$ sous la forme $\hat{p} \mp E$.

Exercice 2.1.2. *Utilisez l'intervalle de confiance fourni pour trouver l'estimation poctuelle $\hat{p}$ et la marge d'erreur E.*

1. $[0,454; 0,494]$

2. $0,642 < p < 0,688$

Exercice 2.1.3. *Une entreprise de production de graines veut vérifier la faculté germinative d'une espèse, c'est à dire la probabilité p pour une graine, prise par hasard dans la production, gèrme.*
Sur un échantillon de 400 graines, on trouve que 330 graines gerrment. Quel est l'intervalle de confiance de p au risque de 5% ? au risque de 1% ?

Exercice 2.1.4. *Pour obtenir une estimation de la proportion d'hyperglycémiques parmi les personnes agées de plus de soixante ans (population Ω), on choisit au hasard 170 personnes dans Ω. On constate que, parmi celles-ci, 31 sont hyperglycémiques.*

1. Donnez un intervalle de confiance à 5% pour le pourcentage p de personnes hyperglycémiques de Ω .

2. Si on effectuait 200 fois le tirage au sort de 170 personnes de Ω. On pourrait construire 200 intervalles de confiance de type précédent. Parmi ces 200 intervalles, combien , en moyenne ,contiendraient la valeur p ?

Exercice 2.1.5. *En acceptant un coefficient de risque $\alpha = 0,05$, on voudrait connaitre à $\pm$ 1 (%) le pourcentage de sujets non immunisés après une certaine vaccination.*
Sur combien de sujets, au minimum, l'observation doit - elle parter ? On sait par avance que le pourcentage d'echecs à cette vaccination est compris entre 10 et 15 %.

Exercice 2.1.6. *On prélève au hasard, dans une population de lapins, 100 individus. Sur ces 100 lapins, 20 sont atteints par la myxomatose. Que peut-on dire du pourcentage des lapins atteints par la myxomatose au niveau de la population ?*