1ère baccalauréat en Sciences Vétérinaires

Biostatistiques – Travaux dirigés

Séance de TD n°3 – Exercices supplémentaires - Résolution

Ex 1

On s'intéresse à la vitesse de course des sous-espèces de guépards : on aimerait déterminer si la sous-espèce asiatique a une vitesse moyenne de course supérieure à celle de la sous-espèce africaine, utilisée comme référence. Cette dernière a une vitesse moyenne de 80 km/h.

Pour répondre à la question, on met le problème sous la forme d'un test d'hypothèse. Sous l'hypothèse nulle (pas de différence de vitesse entre les deux sous-espèces), la vitesse des deux sous-espèces a une distribution normale, de moyenne 80 km/h et de déviation standard 8 km/h. Sous l'hypothèse alternative testée, les guépards asiatiques ont des vitesses distribuées normalement également, mais avec une vitesse moyenne de 82 km/h et la même déviation standard.

Si l'hypothèse alternative est vraie dans 85% des cas, la vitesse de combien de guépards asiatiques (n) faut-il mesurer pour être sûr de détecter la différence entre les deux sous-espèces au seuil $\alpha = 5\%$? Détaillez votre raisonnement pour déterminer la taille d'échantillon satisfaisant aux conditions de l'énoncé.

RESOLUTION

Nous cherchons à déterminer si la vitesse de courses des guépards asiatiques est, en moyenne, supérieure à celle des guépards africains (population de référence).

Dans un premier temps, posons les hypothèses :

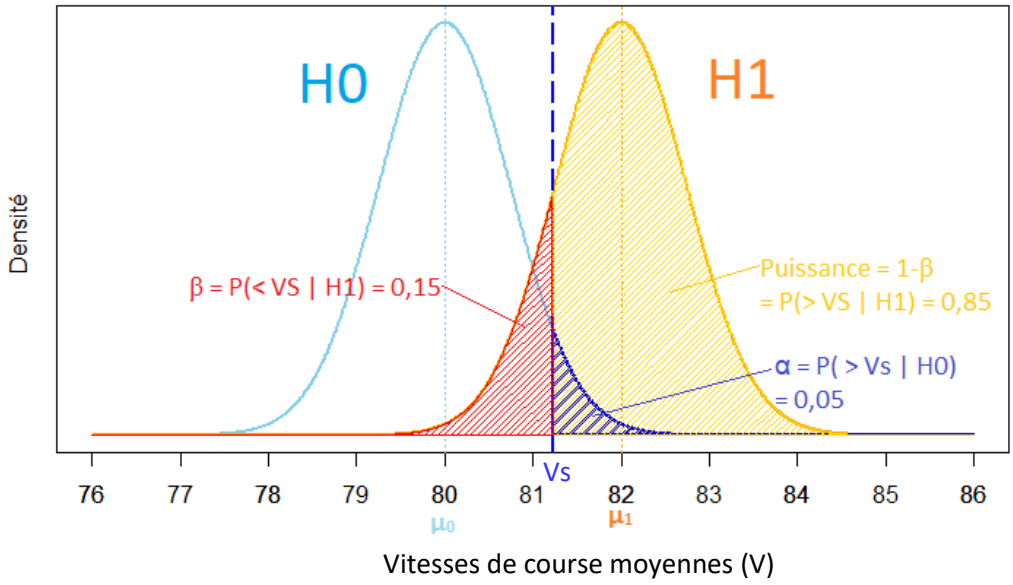
- H0 : la vitesse des guépards asiatiques est en moyenne la même que celle des guépards africains : $\mu_{asiatiques} = \mu_{africains}$
- H1 : la vitesse des guépards asiatiques est en moyenne supérieure à celle des guépards africains : $\mu_{asiatiques} > \mu_{africains}$

Il s'agit donc d'un test unilatéral droit.

Dans cet exercice, on souhaite que :

- Erreur $\alpha = 5\%$. On veut donc que la probabilité $P(V > Vs \mid H0) = 0,05$ (où Vs représente la vitesse seuil, définissant les zones d'acceptation et de rejet de l'hypothèse nulle).
- Puissance = $1 - \beta = 85$ %. On souhaite donc que $P(V > Vs \mid H1) = 0,85$.
- Inversement, erreur $\beta = 1 -$ puissance = 0,15. On souhaite donc que $P(V < Vs \mid H1) = 0,15$

Si on représente graphiquement ces indications, on obtient :



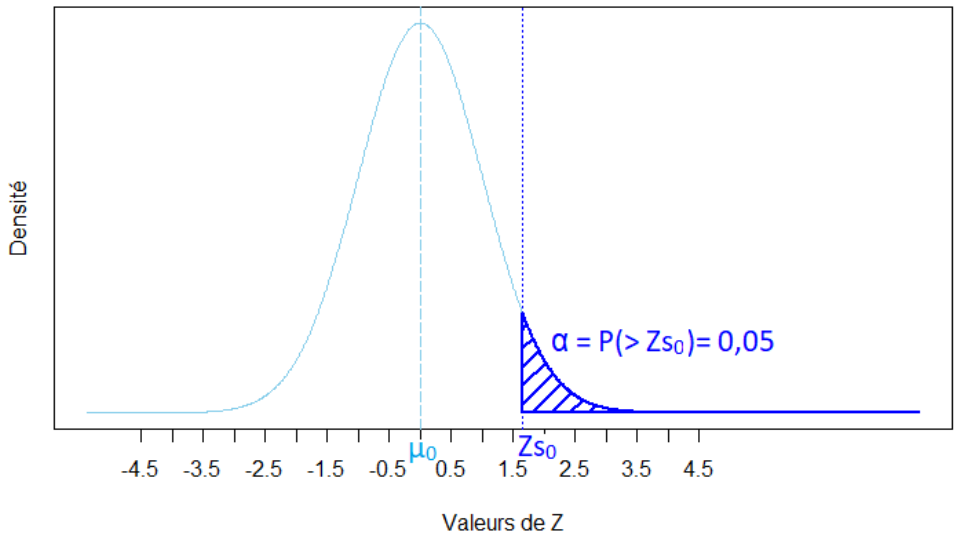
Sachant que, pour un échantillon de taille n, $Z_s = \frac{V_s - \mu}{\frac{\sigma}{\sqrt{n}}}$, nous allons utiliser la loi normale standard afin de déterminer les valeurs Zs correspondantes à chaque distribution. Par la suite, isoler puis égaler les 2 valeurs de Vs obtenues (pour chaque distribution) permettra de déterminer le nombre d'individus n nécessaire pour satisfaire aux conditions.

<u>Comment déterminer les valeurs Zs pour H0 et H1 ?</u>

- Sous H0 : déterminer Zs_0 telle que $P(V > Vs \mid H0) = P(Z > Zs_0) = 0,05$
- Sous H1 : déterminer Zs_1 telle que $P(V > Vs \mid H1) = P(Z > Zs_1) = 0,85$ (ou telle que $P(Z < Zs_1 \mid H1) = 0,15$.

<u>Equation sous H0</u> :

On sait que $P(Z > Zs_0 \mid H0) = 0,05$. On sait déjà que la valeur Zs_0 sera positive.



De fait, en se référant à la table de Z, on détermine :

$P(\mu_0 < Z < Zs_0) = P(Z > \mu_0) - P(Z > Zs_0) = 0{,}5 - 0{,}05 = 0{,}45$

⇨ $Z = 1{,}645$

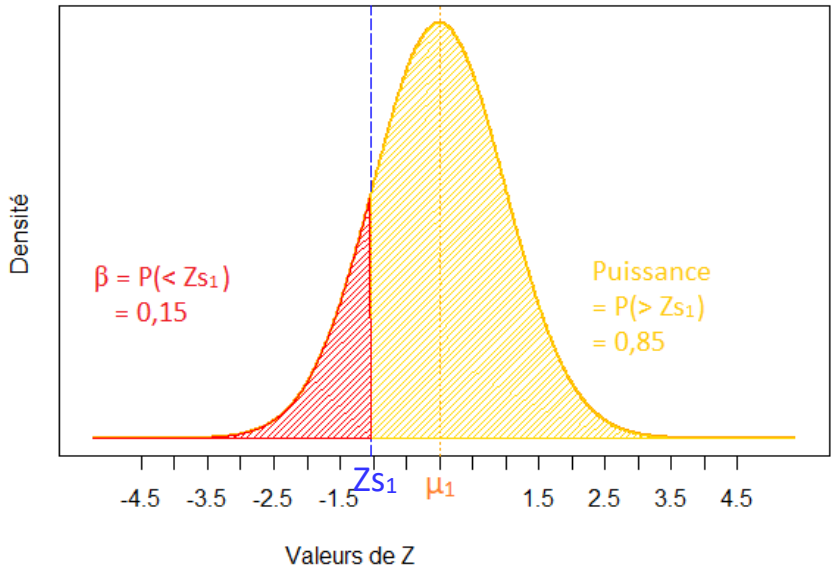
Donc :

$$Z_{s0} = \frac{Vs_0 - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leftrightarrow 1{,}645 = \frac{Vs_0 - 80}{\frac{\sqrt{64}}{\sqrt{n}}} \leftrightarrow 1{,}645 * \frac{\sqrt{64}}{\sqrt{n}} = Vs_0 - 80$$

$$\leftrightarrow Vs_0 = 1{,}645 * \frac{8}{\sqrt{n}} + 80$$

<u>Equation sous H1</u> :

On sait que $P(V > Vs \mid H1) = P(Z > Zs_1) = 0{,}85$ (ou $P(V < Vs \mid H1) = P(Z < Zs_1 \mid H1) = 0{,}15$). On sait déjà que la valeur Zs_1 sera négative.



De fait, en se référant au tableau des Z, on détermine :

$P(Zs_1 < Z < \mu_1) = P(Z < \mu_1) - P(Z < Zs_1) = 0{,}5 - 0{,}15 = 0{,}35$

⇨ $Zs_1 \sim -1{,}04$

Donc :

$$Z_{s0} = \frac{Vs_1 - \mu_1}{\frac{\sigma}{\sqrt{n}}} \leftrightarrow -1{,}04 = \frac{Vs_1 - 82}{\frac{\sqrt{64}}{\sqrt{n}}} \leftrightarrow -1{,}04 * \frac{8}{\sqrt{n}} = Vs_1 - 82$$

$$\leftrightarrow Vs_1 = -1{,}04 * \frac{8}{\sqrt{n}} + 82$$

Egalisation de Vs_0 et Vs_1

$$Vs_0 = Vs_1 \leftrightarrow 1{,}645 * \frac{8}{\sqrt{n}} + 80 = -1{,}04 * \frac{8}{\sqrt{n}} + 82$$

$$\leftrightarrow \frac{8}{\sqrt{n}}(1{,}645 + 1{,}04) = -80 + 95 \leftrightarrow 2{,}649 * \frac{8}{\sqrt{n}} = 15$$

$$\leftrightarrow \sqrt{n} * 2 = 2{,}685 * 8 \leftrightarrow \sqrt{n} = \frac{2{,}685 * 8}{2} \leftrightarrow \sqrt{n} = 10{,}74 \leftrightarrow n = 115{,}3476$$

Il faudrait donc n = 116 individus pour satisfaire aux conditions d'erreurs α, β et de puissance.

Ex 5

Des scientifiques cherchaient à déterminer si les chats préféraient la compagnie de leur maitre à la solitude. Pour cela, ils ont sélectionné 45 chats dans la population, les ont placés dans un couloir menant d'un côté à une pièce vide et de l'autre à une pièce dans laquelle se trouvait leur maitre et ont noté le choix effectué par chaque animal.

Sous l'hypothèse nulle (absence de préférence), il était attendu que les chats choisissent indifféremment une pièce ou l'autre (p_{H0} = 0,5). L'hypothèse alternative était que les chats préfèrent la présence de leur maitre. Ils ont constaté que 27 chats préféraient la pièce où se trouvait leur maitre (soit 60%) mais cette observation n'a pas permis de démontrer de différence significative par rapport à ce qui était attendu (prouvez le !). Ils se demandent cependant si cette absence de différence significative n'est pas due à un manque de puissance statistique de leur expérience, si effectivement, en réalité, 6 chats sur 10 préfèrent leur maitre à une pièce vide.

Sachant qu'on considère une erreur α = 5%, quelle est la puissance de cette expérience ?

RESOLUTION

Nous sommes face à une variable discrète binaire (pièce vide *vs* pièce avec le propriétaire), avec n = 45 tirages et des tirages « avec remise » (à l'échelle de la population de chats)

⇨ Loi binomiale

Rappelons les hypothèses :

- H0 : les chats n'ont pas de préférence pour une pièce ou l'autre : $p_{propriétaire} = p_{vide} = 0{,}5$
- H1 : les chats préfèrent la présence de leur propriétaire : $p_{propriétaire} > p_{vide}$

Test de l'hypothèse

Pour démontrer une préférence ou absence de préférence des chats pour leur maitre, on peut calculer $P(r \geq 27 \mid H0)$, c'est-à-dire $P(r \geq 27)$ sous la loi binomiale de caractéristiques B(n = 45 ; p = 0,5). On

calcule ainsi la probabilité d'observer cette situation et les situations qui s'en éloignent davantage sous l'hypothèse nulle, ce qui correspond à la p-value.

On obtient les probabilités suivantes avec la loi binomiale :

r	P(=r | H0)
27	0,0487684
28	0,0313511
29	0,0183782
30	0,0098017
31	0,0047428
32	0,002075
33	0,0008174
34	0,0002885
35	9,067E-05
36	2,519E-05
37	6,126E-06
38	1,29E-06
39	2,315E-07
40	3,472E-08
41	4,235E-09
42	4,033E-10
43	2,814E-11
44	1,279E-12
45	2,842E-14
SOMME	0,1163466

$P(r \geq 27 \mid H0) = 0,1163466 > 0,05$ donc on accepte l'hypothèse nulle : il semble que les chats ne marquent pas de préférence pour leur maitre.

Recourir au calcul exact (ici avec la loi binomiale) peut-être long. Il est aussi possible dans notre situation d'utiliser un Khi² de conformité afin de déterminer si la répartition observée est conforme à la répartition attendue :

	Observés	Attendus
Pièce avec propriétaire	27	45*0,5 = 22,5
Pièce vide	45 – 27 = 18	22,5

$$\chi^2_{obs} = \sum \frac{(obs-att)^2}{att} = \frac{(27-22,5)^2}{22,5} + \frac{(18-22,5)^2}{22,5} = 1,8$$

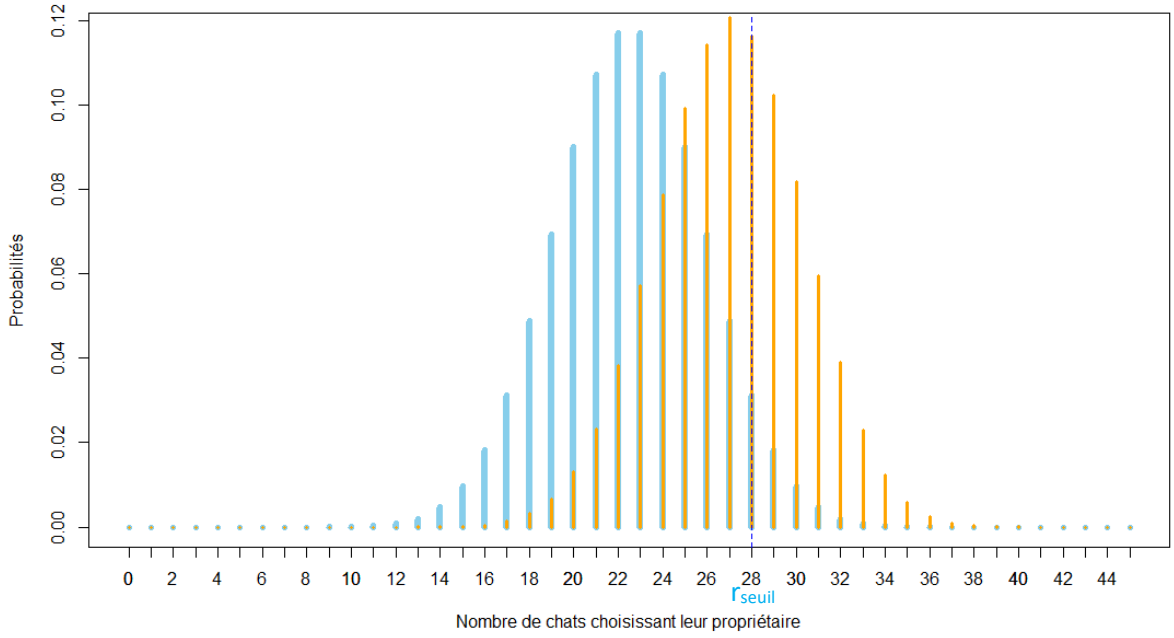
$\chi^2_{seuil\ 5\%}$ pour 1 ddl = 3,841

$\chi^2_{obs} < \chi^2_{seuil\ 5\%}$ donc on accepte H0 : la distribution observée semble conforme à ce qui est attendu. Les chats n'auraient donc pas de préférence.

On peut calculer la p-value avec un logiciel : $P(>\chi^2_{obs}) = 0{,}179712 > 0{,}05$. Cette probabilité est similaire à celle obtenue à l'aide de la loi binomiale et mène également à l'acceptation de H0.

Détermination de la puissance

Nous pouvons commencer par représenter graphiquement les 2 distributions :



En bleu la distribution H0, en orange la distribution H1

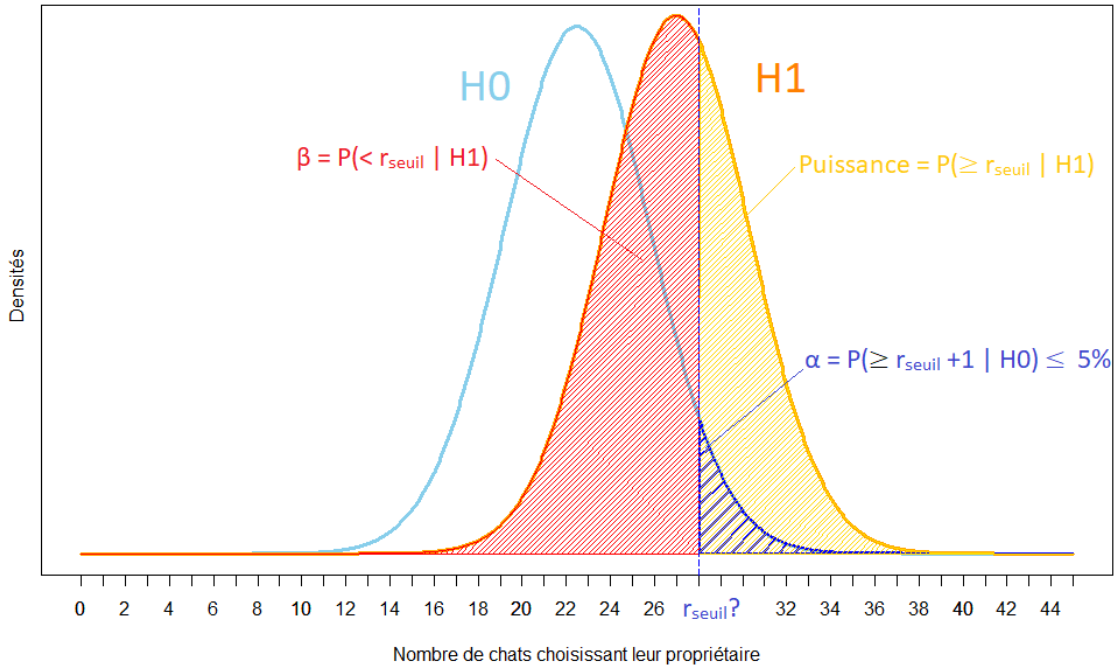
Dans cet exercice, on sait que :

- Erreur $\alpha = 5\%$.
- $p_{H1} = 0{,}6 > p_{H0} = 0{,}5$ donc le test est unilatéral droit
- ⇨ Ces informations vont nous permettre de déterminer la valeur r_{seuil}. Remarque : pour les distributions discrètes, il est rare d'obtenir une valeur seuil correspondant exactement à 5%. On veut donc que l'erreur α soit la plus proche possible de 5%, tout en y restant inférieure et on définit la valeur r_{seuil} telle que $P(\geq r_{seuil} + 1 \mid H0) \leq 5\%$ et $P(\geq r_{seuil} \mid H0) > 5\%$. En somme, $\alpha = P(\geq r_{seuil} + 1 \mid H0)$.

Une fois la valeur r_{seuil} déterminée, on peut calculer la puissance et l'erreur β, telles que :

- $\beta = P(< r_{seuil} \mid H1)$
- Puissance = $1 - \beta = P(\geq r_{seuil} \mid H1)$

Nous pouvons représenter ces données sous la forme d'une courbe de Gauss afin d'illustrer plus facilement les différentes probabilités cumulées :



Réaliser les calculs exacts (avec la loi binomiale) est très rapide à l'aide d'un logiciel mais il serait long de déterminer manuellement r_{seuil} et de calculer les différentes probabilités avec la loi binomiale. Or, sachant que n*p > 5 et n*(1-p) > 5, nous pouvons utiliser l'approximation normale.

Nous devons donc dans un premier temps déterminer Z_{seuil} sous H0 (notée Zs_0), telle que P(> Zs_0) = 5%* puis en déduire la valeur r_{seuil} correspondant sous la loi binomiale B(n=45 ; p_{H0} = 0,5). *<u>Remarque</u> : comme on passe par une distribution continue, on peut déterminer Zs_0 de sorte que P(> Zs_0) = 5%.

Pour déterminer la puissance et l'erreur β, nous devrons une fois de plus passer par l'approximation normale. Déterminer la valeur Z_{seuil} sous H1 (notée Zs_1) à partir de r_{seuil} nous permettra alors de calculer la puissance et l'erreur β telles que :

- β = P(< Zs_1)
- Puissance = 1- β = P(> Zs_1)

<u>Détermination de la valeur seuil (sous H0)</u>

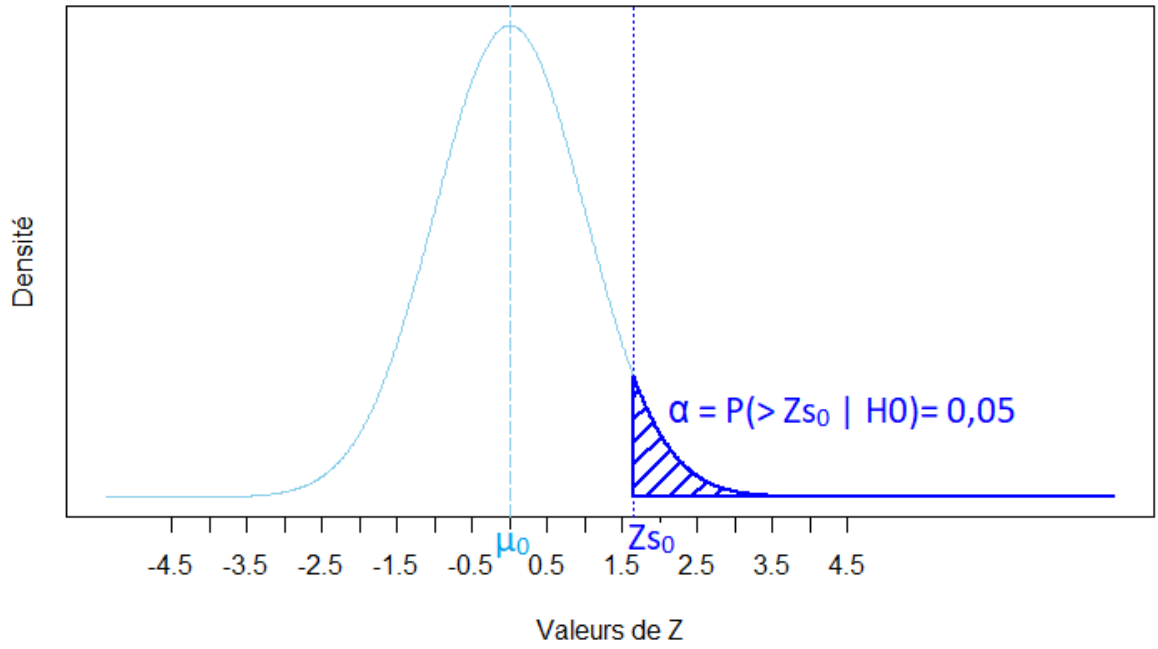
On sait que P(Z > Zs_0) = 0,05 donc P(0 < Z < Zs_0) = P(Z > 0) - P(Z > Zs_0) = 0,5 – 0,05 = 0,45

⇨ Zs_0 = 1,645

Dès lors, sachant que $Zs_0 = \frac{r_{seuil} - n*p_{H0}}{\sqrt{n*p_{H0}*(1-p_{H0})}}$,

alors $r_{seuil\ H0} = Zs_0 * \sqrt{n * p_{H0} * (1 - p_{H0})} + n * p_{H0}$

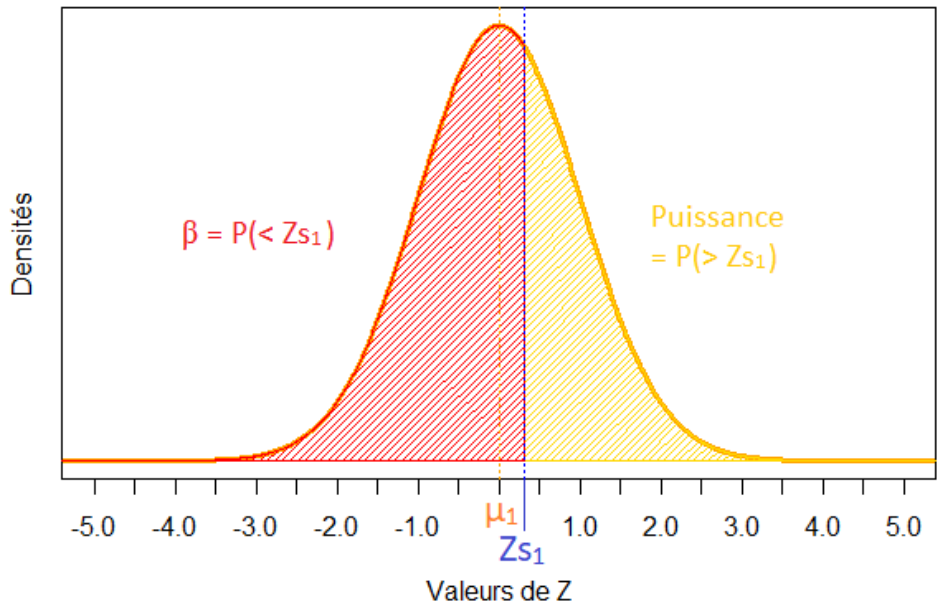
$= 1{,}645 * \sqrt{45 * 0{,}5 * 0{,}5} + 45 * 0{,}5 = 28{,}017498$



Détermination de la puissance et de l'erreur β (sous H1)

On calcule $Zs_1 = \frac{r_{seuil} - n*p_{H1}}{\sqrt{n*p_{H1}*(1-p_{H1})}} = \frac{28{,}017498 - 45*0{,}6}{\sqrt{45*0{,}6*0{,}4}} = 0{,}309615$

⇨ $Zs_1 \sim 0{,}31$ (arrondir pour lire la table de Z).

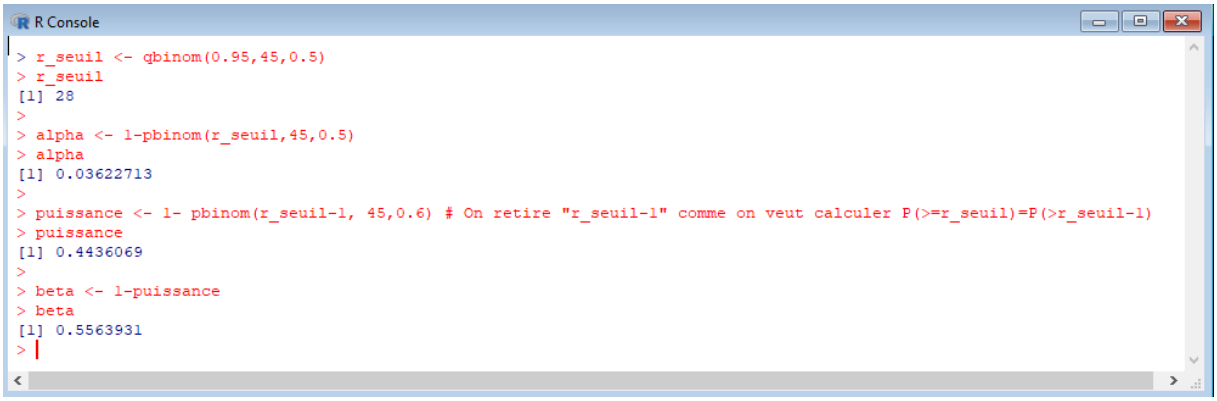


Nous pouvons dès lors déterminer puissance = P(> Zs_1) = P(Z > 0) - P(0 < Z < Zs_1) = P(Z > 0) - P(0 < Z < 0,31) = 0,5 – 0,1217 = 0,3783 et β = 1 – puissance = 0,6217.

La puissance est donc très faible (~38%) et l'erreur β inversement très importante (~62%). Il est donc tout à fait possible que les chercheurs aient manqué un effet qui existe du fait de la faible puissance de leur étude. Remarque : on attend usuellement une puissance d'au moins 80% dans les études scientifiques. Plus la taille d'échantillon est importante et/ou l'effet recherché important et plus la puissance augmente.

Pour info :

Comme indiqué précédemment, calculer la valeur seuil, α, β et la puissance est très facile à l'aide d'un logiciel. Par exemple dans R, on peut les obtenir de la manière suivante (ce que vous allez notamment apprendre en TP) :



On remarque que la puissance est de 44%, soit une puissance plus élevée que ce que nous avons trouvé avec l'approximation normale : l'approximation a donc mené à une sous-estimation de la puissance (qui reste malgré tout bien trop faible).