Indications sur l'activité 1.

1.1. Tirages possibles (sans remplacement) de numéros dans les deux populations (avant, après) :

Avant	14	2	6	9	19
Après	8	1	13	15	7

Tensions observées :

Avant :

i	$x_{1,i}$
1	16
2	13
3	13
4	15
5	17
$\overline{x_1}$	14,8

Après :

i	$x_{2,i}$
1	15
2	14
3	11
4	13
5	15
$\overline{x_2}$	13,5

Au vu de ces deux séries, puisque $\overline{x_1} = 14{,}8 > 13{,}5 = \overline{x_2}$, les tenants de la méthode numérique auraient conclu à l'efficacité du médicament.

Mais il y a d'autres tirages possibles, par exemple :

Avant	2	4	18	20	14
Après	19	3	7	1	12

Tensions observées :

Avant :

i	$x_{1,i}$
1	13
2	11
3	12
4	13
5	16
$\overline{x_1}$	13

Après :

i	$x_{2,i}$
1	15
2	18
3	15
4	14
5	15
$\overline{x_2}$	15,4

Au vu de ces deux nouvelles séries, puisque $\overline{x_1} = 13 < 15{,}4 = \overline{x_2}$, les tenants de la méthode numérique n'auraient pas conclu à l'efficacité du médicament.

1.2. Moyennes effectives avant et après : $\mu_1 = 14$ et $\mu_2 = 14$.

Écarts des tensions à la moyenne avant traitement :

α	1	2	3	4	5	6	7	8	9	10
$\epsilon_{1,\alpha}$	-2	-1	2	-3	0	-1	1	2	1	-2
α	11	12	13	14	15	16	17	18	19	20
$\epsilon_{1,\alpha}$	1	1	-2	2	1	0	0	-2	3	-1

Écarts des tensions à la moyenne après traitement :

α	1	2	3	4	5	6	7	8	9	10
$\epsilon_{2,\alpha}$	0	0	4	-2	-2	-2	1	1	-1	-2
α	11	12	13	14	15	16	17	18	19	20
$\epsilon_{2,\alpha}$	2	1	-3	4	-1	2	0	-3	1	0

Il est aisé de vérifier que : $\sum_{\alpha=1}^{20} \epsilon_{1,\alpha} = 0$ et $\sum_{\alpha=1}^{20} \epsilon_{2,\alpha} = 0$.

Ainsi donc, si on entend par “erreur” l'écart de la valeur de la tension à la moyenne des valeurs de **toute** la population, les “erreurs” se compensent. On pouvait prévoir le résultat car :

$\sum_{\alpha=1}^{20} \epsilon_{1,\alpha,} = \sum_{\alpha=1}^{20}(\xi_{1,\alpha} - \mu_1) = (\sum_{\alpha=1}^{20} \xi_{1,\alpha}) - 20\mu_1 = 20\mu_1 - 20\mu_1 = 0.$

Même raisonnement pour les écarts calculés après traitement. L'assertion de Louis semble donc vérifiée.

1.3. Dans le cas de la première série de tirages :

Avant :

i	$e_{1,i}$
1	2
2	-1
3	-1
4	1
5	3
$\sum_{i=1}^{5} e_{1,i}$	4

Après :

i	$e_{2,i}$
1	1
2	0
3	-3
4	-1
5	2
$\sum_{i=1}^{5} e_{2,i}$	-1

Ici, les “erreurs” ne se compensent pas.

Mais par exemple, pour les tirages suivants :

Avant	8	15	4	15	2
Après	10	3	4	19	9

les “erreurs” se compensent.

Gavarret avait donc raison en émettant des doutes sur le fait que les "erreurs" puissent se compenser. Cette compensation est vraie si c'est l'ensemble de la population qui est prise en compte mais elle ne l'est plus obligatoirement si c'est seulement une partie de la population étudiée qui est examinée. En effet, $\sum_{\alpha=1}^{20} \epsilon_{1,\alpha} = 0$ n'implique pas obligatoirement $\sum_{\alpha \in A} \epsilon_{1,\alpha} = 0$ pour $A \subset \{1, 2, ..., 19, 20\}$.

Si les “erreurs” se compensent, alors $\overline{x_1} - \overline{x_2} = (\mu_1 + \sum_{i=1}^{5} e_{1,i}) - (\mu_2 + \sum_{i=1}^{5} e_{2,i}) = \mu_1 - \mu_2$ et $\overline{x_1} > \overline{x_2}$ entraîne $\mu_1 > \mu_2$. Gavarret a bien compris qu'à partir où les observations se font sur une partie seulement de la population, la possibilité d'étendre les conclusions issues de ces observations à la population toute entière n'allait pas de soi.

1.4. Il semble plus pertinent de travailler sur les différences entre la tension après et la tension avant et de procéder ainsi :

Le médecin tire au hasard 5 numéros parmi les 20. Pour chacun des patients, il examine la tension avant et après. En effet, ce type d'expérimentation tiendra davantage compte de l'effet “patient” sur le médicament. Si un patient a une tension très élevée avant la prise du médicament et que celui-ci a un effet, la tension restera cependant élevée mais c'est la différence qu'il faudra prendre en compte.

Commentaire : Le but de cette activité est de faire prendre conscience de la différence entre la population étudiée et l'échantillon observé, et de la non-évidence à étendre à toute la population des conclusions obtenues à la suite d' observations issues d' un échantillon.

Indications sur l'activité 2.1

2.1.1. Exemple de possibilité d'obtention d'un tirage de 900 **blanches et** 100 **noires avec** $K < H$**.**

Si le tirage a lieu **avec remplacement**, il suffit de prendre N quelconque ($N > 2$ et $K < \frac{N}{2}$).
Par exemple, $N = 3$, $K = 1$ et $H = 2$. Ou encore $N = 1000$, $K = 499$ et $H = 501$.

Si le tirage a lieu **sans remplacement**, il faut prendre $K \geq 900$ et $N > 2K$.
Par exemple $N = 1801$, $K = 900$ et $H = 901$ conviennent mais aussi $N = 2000$, $K = 999$ et $H = 1001$, ou $N = 2000$, $K = 950$ et $H = 1050$.

2.1.2. Exemple de possibilité d'obtention d'un tirage de 900 **blanches et** 100 **noires avec** $K = H$**.**

Si le tirage a lieu **avec remplacement**, il suffit de prendre K quelconque et $N = 2K$.
Par exemple, $N = 2$, $K = 1$ et $H = 1$. Ou encore $N = 1000$, $K = 500$ et $H = 500$ mais aussi $N = 2000$, $K = 1000$ et $H = 1000$.

Si le tirage a lieu **sans remplacement**, il faut prendre $K \geq 900$ et $N = 2K$.
Par exemple, $N = 1800$, $K = 900$ et $H = 900$ conviennent.

> *Commentaire :* Il s'agit, par des exemples, de montrer que l'observation d'un beaucoup plus grand nombre de blanches que de noires n'implique pas qu'il y ait plus de blanches que de noires dans l'urne.

Indications sur l'activité 2.2

2.2.1. En réalité, Gavarret énonce la conclusion $K > H$ à partir de l'observation suivant laquelle le nombre de boules blanches, dans un tirage de 1000 boules, est très supérieur à un nombre donné (ici il propose 900). On voit mal comment il conclurait $K > H$ en observant 900 blanches mais conclurait le contraire en observant 901 blanches, ..., ou 902 blanches ! Pour choisir entre les deux hypothèses : $K \leq H$ et $K > H$, la règle de décision est implicitement la suivante : "Soit X la variable qui associe à un tirage de 1000 boules le nombre x de blanches. On décide $K > H$ si $x \geq A$, où A est un nombre tel que, si en réalité on avait $K \leq H$ la probabilité de l'événement $(X \geq A)$ serait très petite" (ici, Gavarret prend $A = 900$).

En effet, dans **2.1.**, il est montré que, bien que x soit supérieur ou égal à 900, on ne peut exclure que $K \leq H$. Il faut seulement, dans ce cas, que la probabilité de se tromper en décidant $K > H$ soit très faible. Vérifions-le sur quelques exemples, où $K \leq H$.

2.2.2. Dans le cas d'un tirage **avec remplacement**, X suit une loi binomiale de paramètres $n = 1000$ et $p = \dfrac{K}{N}$.

$$\mathbb{P}(X = x) = \binom{1000}{x} p^x (1-p)^{1000-x}$$

Dans le cas d'un tirage **sans remplacement**, X suit une loi hypergéométrique :

$$\mathbb{P}(X = x) = \frac{\dbinom{K}{x}\dbinom{H}{1000-x}}{\dbinom{N}{1000}}$$ pour $0 \leq x \leq K$ et $0 \leq 1000 - x \leq H$, et $\mathbb{P}(X = x) = 0$ sinon.

- **Cas avec remplacement et** $N = 3$**,** $K = 1$ **et** $H = 2$**.**
  X suit une loi binomiale de paramètres $n = 1000$ et $p = \frac{1}{3}$. On utilise l'approximation normale c'est-à-dire $X \sim \mathcal{N}(np, \sqrt{np(1-p)})$.
  $\mathbb{P}(X \geq 900) = 1 - \mathbb{P}(X < 900) = 1 - \mathbb{P}(X \leq 899) = 1 - \mathbb{P}\left(\frac{X - \frac{1000}{3}}{\sqrt{1000 \times \frac{1}{3} \times \frac{2}{3}}} \leq \frac{899 - \frac{1000}{3}}{\sqrt{1000 \times \frac{1}{3} \times \frac{2}{3}}}\right)$
  = $1 - \mathbb{P}(Z \leq 37{,}94) = 1 - F_Z(37{,}94)$ où F_Z est la fonction de répartition de la loi normale centrée réduite. Or $F_Z(37{,}94) \simeq 1$ donc $\mathbb{P}(X \geq 900) \simeq 0$

- **Cas avec remplacement et** $N = 1000$**,** $K = 499$ **et** $H = 501$**.**
  X suit une loi binomiale de paramètres $n = 1000$ et $p = \frac{499}{1000}$.
  Le même raisonnement montre que : $\mathbb{P}(X \geq 900) = 1 - F_Z(25{,}09) \simeq 0$.

- **Cas avec remplacement et** $N = 2$**,** $K = 1$ **et** $H = 1$**.** Ou encore $N = 1000$, $K = 500$ et $H = 500$ mais aussi $N = 2000$, $K = 1000$ et $H = 1000$.
  X suit une loi binomiale de paramètres $n = 1000$ et $p = 0{,}50$.
  Le même raisonnement montre que $\mathbb{P}(X \geq 900) = 1 - \mathbb{P}(Z \leq 25{,}23) \simeq 0$.

- **Cas sans remplacement avec** $N = 1801$**,** $K = 900$ **et** $H = 901$**.**
  Ici $\mathbb{P}(X \geq 900) = \mathbb{P}(X = 900) = \dfrac{\binom{900}{900}\binom{901}{100}}{\binom{1801}{1000}} \simeq 0$.

- **Cas sans remplacement avec** $N = 2000$**,** $K = 999$ **et** $H = 1001$. Ou $N = 2000$, $K = 950$ et $H = 1050$, ou encore $N = 1800$, $K = 900$ et $H = 900$.
  On a toujours : $\mathbb{P}(X \geq 900) \simeq 0$.

Commentaire : La question 2.2.1. vise à faire comprendre que dans un test où on souhaite décider entre $H_0 : K \leq H$ et $H_1 : K > H$, après l'observation de x blanches, la zone de rejet de l'hypothèse nulle est de la forme $X \geq A$ et non $X = A$, où A est un nombre "assez grand". Même si ici la forme de la zone de rejet paraît évidente, la détermination de celle-ci dans d'autres problèmes de test reste un sujet mathématiquement délicat. La question 2.2.2. permet de faire travailler sur les tables de la fonction de répartition de la loi normale centrée réduite et de montrer, en liaison avec le 2.1., qu'un événement de probabilité nulle n'est pas un événement impossible.

Indications sur l'activité 2.3

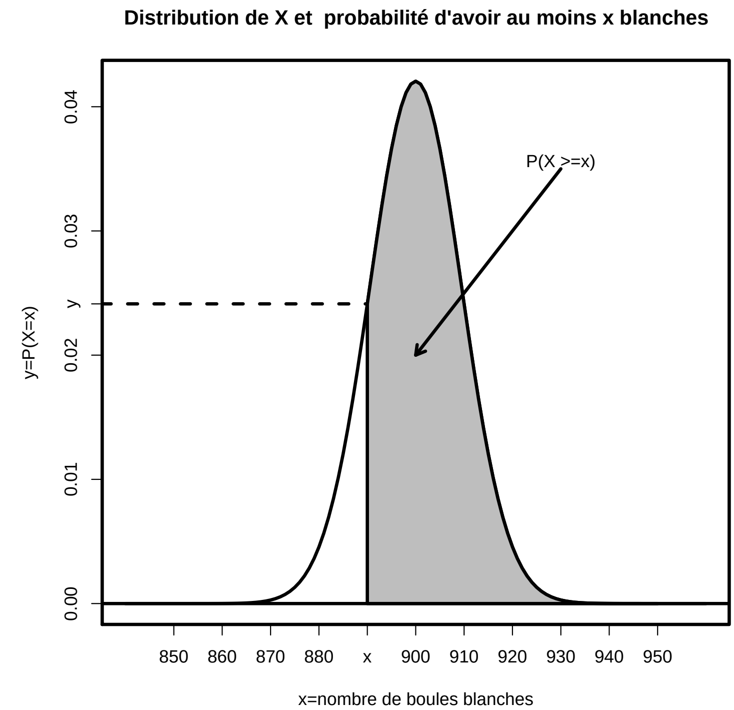
a) Soit X la variable aléatoire qui donne le nombre de blanches en supposant que la proportion de boules blanches dans l'urne est 90%. Alors X suit une loi binomiale de paramètre $n_2 = 1000$ et $p = 0{,}90$. Obtenir plus de blanches que de noires se traduit par la réalisation de $(X > 500)$ ou $(X \geq 501)$.

En utilisant l'approximation normale, on obtient :

$$\mathbb{P}(X \geq 501) = 1 - F_Z\left(\frac{500 - 1000 \times 0{,}90}{\sqrt{1000 \times 0{,}90 \times 0{,}10}}\right) = 1 - F_Z(-4{,}216) = 1 - 0{,}000\,012 = 0{,}999\,988.$$

Cette probabilité très proche de 1 semble donner raison à Gavarret.

b) Si la proportion de boules blanches dans l'urne est 90%, on a $\mathbb{P}(X \geq 950) \simeq 0$! ! ! ! Donc une probabilité presque nulle d'avoir "beaucoup plus de blanches que de noires" (si on considère que "plus de 950 " est équivalent à "beaucoup plus de blanches que de noires"). Ce paradoxe s'explique par le fait que, si la proportion de boules blanches dans l'urne est 90%, la distribution de X se concentre autour de son espérance c'est-à-dire 900 et la probabilité d'obtenir des valeurs éloignées de cette espérance est faible.



Commentaire : Puisque dans la suite de son texte, Gavarret semble considérer comme "absolument certain" [G, p. 257] tout événement dont la probabilité est supérieure ou égale à $0{,}9953$, Gavarret aurait raison d'affirmer qu'on obtiendrait plus de blanches que de noires. Mais peut-il dire qu' "on obtiendrait **beaucoup** plus de blanches que de noires" (c'est nous qui soulignons) ? C'est toute l'ambiguïté du mot "beaucoup" et qui se retrouvera plus loin dans le texte, et qui peut donner des résultats apparemment paradoxaux comme le montrera 2.4.5.

Indications sur l'activité 2.4

2.4.1. a) Déjà montré en **2.2.2.** : $\mathbb{P}(X \geq 900) = 1 - \mathbb{P}(Z \leq 25{,}23) \simeq 0 < 0{,}0047$, donc ce cas est "impossible au sens de Gavarret".

Calcul de la probabilité des événements suivants :

– "tirer au moins 546 blanches". $\mathbb{P}(X \geq 546) = 1 - F_Z\left(\dfrac{545 - 1000 \times 0{,}50}{\sqrt{1000 \times 0{,}50 \times 0{,}50}}\right) =$

$= 1 - \mathbb{P}(Z \leq 2{,}846) = 0{,}002\,213 < 0{,}0047$. Événement "impossible au sens de Gavarret" ;

– "tirer au moins 543 blanches". $\mathbb{P}(X \geq 543) = 1 - \mathbb{P}(Z \leq 2{,}656) = 0{,}00395 < 0{,}0047$.
Événement impossible "impossible au sens de Gavarret" ;

– "tirer au moins 542 blanches". $\mathbb{P}(X \geq 542) = 1 - \mathbb{P}(Z \leq 2{,}593) = 0{,}00475 > 0{,}0047$;

– " tirer au moins 541 blanches". $\mathbb{P}(X \geq 541) = 1 - \mathbb{P}(Z \leq 2{,}530) = 0{,}00570 > 0{,}0047$.

b) Il est aisé de démontrer que la fonction u définie ainsi :
$u(x) = \dfrac{(x-1) - 1000 \times 0{,}50}{\sqrt{1000 \times 0{,}50 \times 0{,}50}}$ est strictement croissante et puisque la fonction F_Z est aussi strictement croissante, la fonction g définie ainsi :
$g(x) = \mathbb{P}(X \geq x) = 1 - F_Z(u(x))$ est alors strictement décroissante.

Remarque : L'utilisation des tables montre que $F_Z(2{,}597) = 0{,}9953$ et donc que si Z est une loi normale centrée réduite, alors $\mathbb{P}(Z \geq 2{,}597) = 0{,}0047$. On peut trouver A en cherchant le plus petit entier x tel que $(x-1) - 1000 \times 0{,}50 \geq 2{,}597\sqrt{1000 \times 0{,}50 \times 0{,}50}$. Puisque $501 + 2{,}597\sqrt{1000 \times 0{,}50 \times 0{,}50} = 542{,}062$, on trouve bien $A = 543$.

Ainsi donc, si dans l'urne il y a autant de blanches que de noires, $A = 543$ est le plus petit des entiers x tels que, l'événement "on a tiré au moins x boules blanches à la suite de 1000 tirages" est impossible "au sens de Gavarret".

Commentaire : Ici, après avoir en 2.2.1. déterminé la forme : $X \geq A$ de la zone de rejet de l'hypothèse nulle, il s'agit de déterminer A de telle sorte que, quand si l'hypothèse nulle est vraie, alors la probabilité que le résultat de l'expérience tombe dans cette zone de rejet soit inférieure à $0{,}0047$. Ici le calcul explicite de A s'effectue en supposant que la composition de l'urne est telle que $K = H$, c'est-à-dire quand $p = \dfrac{K}{N} = 0{,}50$. Pour être rigoureux, il faudrait prouver aussi que la probabilité que le résultat de l'expérience tombe dans cette zone de rejet est inférieure à $0{,}0047$, non seulement quand $K = H$, mais aussi quand $K < H$ ("moins de blanches que de noires"). C'est l'objet de la question suivante.

2.4.2. a) Si la proportion de blanches dans l'urne est 49%, la probabilité de tirer au moins 543 blanches est égale à :

$$\mathbb{P}(X \leq 543) = 1 - F_Z\left(\frac{542 - 1000 \times 0{,}49}{\sqrt{1000 \times 0{,}49 \times (1 - 0{,}49)}}\right) = 1 - \mathbb{P}(Z \leq 3{,}289) = 0{,}000\,502 < 0{,}0047.$$

Si la proportion de blanches dans l'urne est 48%, la probabilité de tirer au moins 543 blanches est égale à :

$$\mathbb{P}(X \leq 543) = 1 - F_Z\left(\frac{542 - 1000 \times 0{,}48}{\sqrt{1000 \times 0{,}48 \times (1 - 0{,}48)}}\right) = 1 - \mathbb{P}(Z \leq 3{,}924) = 0{,}000\,043 < 0{,}0047.$$

$\mathbb{P}_{0{,}48}(X \geq 543) \leq \mathbb{P}_{0{,}49}(X \geq 543) \leq \mathbb{P}_{0{,}50}(X \geq 543) < 0{,}0047$.

b) On pose h la fonction définie par :
$h(p) = \mathbb{P}_p(X \geq 543) = 1 - F_Z\Big(\dfrac{542 - 1000 \times p}{\sqrt{1000 \times p \times (1-p)}}\Big)$.
Calcul de la dérivée :
$$h^{'}(p) = \frac{1000(542 - p(2 \times 542 - 1000))}{2 \times (1000p(1-p))^{\frac{3}{2}}} \frac{1}{\sqrt{2\pi}} e^{-\dfrac{(542-1000p)^2}{2 \times 1000p(1-p)}}$$
or $542 - p(2 \times 542 - 1000) > 0$ car $p < 1 < \dfrac{542}{2 \times 542 - 1000} = 6{,}452$, donc $h^{'}(p) > 0$ pour $0 < p < 1$ et la fonction h est croissante.

Remarque : Cette propriété est vraie pour tout A tel que $\dfrac{n}{2} \leq A \leq n$ car $\dfrac{A}{2A - n} \geq 1 \geq p$.

On peut donc généraliser.

Soit $0 < \alpha < 1$. On pose $\mathbb{P}_p(X \geq x)$ la probabilité d'avoir au moins x boules blanches parmi n tirées dans une urne contenant une proportion p de blanches et A le plus petit entier x tel que $\mathbb{P}_{0.50}(X \geq x) \leq \alpha$ alors si c'est l'hypothèse $K < H$ qui est vraie, on a donc encore $\mathbb{P}_p(X \geq A) < \mathbb{P}_{0{,}50}(X \geq A) \leq \alpha$.

c) Dans l'exemple de Gavarret, si l'hypothèse $K < H$ est vérifiée, $p < 0{,}50$ donc $P_p(X \geq 543) < P_{0{,}50}(X \geq 543) = 0{,}00395 < 0{,}0047$.

Commentaire : Le problème de test peut se ramener ici à celui-ci : choisir entre l'hypothèse nulle $H_0 : p \leq 0{,}50$ et l'hypothèse alternative $H_1 : p > 0{,}50$, avec un risque de se tromper $\alpha = 0{,}0047$ en décidant H_1 et un échantillon de taille $n = 1000$. On vient de montrer que pour calculer explicitement A, il suffisait de trouver le plus petit entier x tel que $\mathbb{P}_p(X \geq x) \leq \alpha$, mais en se restreignant à $p = 0{,}50$ c'est-à-dire la “frontière” entre les deux zones possibles pour le paramètre inconnu.

2.4.3. Montrons que les deux règles de décision sont équivalentes.

Si $x \geq 543$ alors $(X \geq x) \subset (X \geq 543)$ donc $\mathbb{P}_{0{,}50}(X \geq x) \leq \mathbb{P}_{0{,}50}(X \geq 543)$.

Or 543 a été choisi tel que $\mathbb{P}_{0{,}50}(X \geq 543) \leq \alpha$, donc $\mathbb{P}_{0{,}50}(X \geq x) \leq \alpha$.

Si $\mathbb{P}_{0{,}50}(X \geq x) \leq \alpha$, $x < 543$ est impossible car, par construction 543 est le plus petit entier x tel que $\mathbb{P}_{0{,}50}(X \geq x) \leq \alpha$. Donc $x \geq 543$.

Commentaire : Aujourd'hui, tous les logiciels et les calculettes utilisent la notion de **p-value** notée souvent **P** dans les calculettes. Ici, il s'agit de **p-value = P =** $\mathbb{P}_{0{,}50}(X \geq x)$ où x est le résultat (ici le nombre de blanches) obtenu après l'expérience. Pour conclure un test, il suffira de comparer cette **p-value** au risque de première espèce choisi α. Si **p-value** $\leq \alpha$, on rejette l'hypothèse nulle et si **p-value** $> \alpha$, on ne rejette pas l'hypothèse nulle.

Il est maintenant d'usage d'accompagner les conclusions d'un test de sa “signification” suivant les valeurs de la **p-value**.

	Résultat	Notation
$0{,}05 <$ **p-value**	“pas significatif”	
$0{,}01 \leq$ **p-value** $< 0{,}05$	“significatif”	*
$0{,}001 \leq$ **p-value** $< 0{,}01$	“très significatif”	**
p-value $< 0{,}001$	“hautement significatif”	***

2.4.4.

a) Calcul de $\beta(p) = \mathbb{P}_p(X < 543) = \mathbb{P}_p(X \leq 542)$ pour :

- Pour $p = 0{,}51$, $\beta(p) = \mathbb{P}_p(X < 543) = 0{,}978\,528$;
- Pour $p = 0{,}52$, $\beta(p) = \mathbb{P}_p(X < 543) = 0{,}918\,117$;
- Pour $p = 0{,}53$, $\beta(p) = \mathbb{P}_p(X < 543) = 0{,}776\,467$;
- Pour $p = 0{,}54$, $\beta(p) = \mathbb{P}_p(X < 543) = 0{,}550\,489$;
- Pour $p = 0{,}55$, $\beta(p) = \mathbb{P}_p(X < 543) = 0{,}305\,546$;
- Pour $p = 0{,}60$, $\beta(p) = \mathbb{P}_p(X < 543) = 0{,}000\,090$;

$\beta(p) = 1 - h(p)$ or h est croissante donc β est décroissante.

b) **Calcul avec un risque** $\alpha = 0{,}05$. Cas classique.

D'après 2.4.2., pour trouver A^* il suffit de trouver le plus petit x tel que $\mathbb{P}_{0,50}(X \geq x) < 0{,}05$ ou encore le plus petit x tel que $P_{0,50}(X \leq x - 1) > 0{,}95$.

$$\mathbb{P}_{0,50}(X \leq x-1) = \mathbb{P}_{0,50}\left(\frac{X - 1000 \times 0.50}{\sqrt{1000 \times 0{,}50 \times 0{,}50}} \leq \frac{(x-1) - 1000 \times 0{,}50}{\sqrt{1000 \times 0{,}50 \times 0{,}50}}\right) = F_Z\left(\frac{(x-1) - 1000 \times 0{,}50}{\sqrt{1000 \times 0{,}50 \times 0{,}50}}\right)$$

En utilisant une table ou une calculette, on trouve $F_Z(1{,}65) = 0{,}95$.

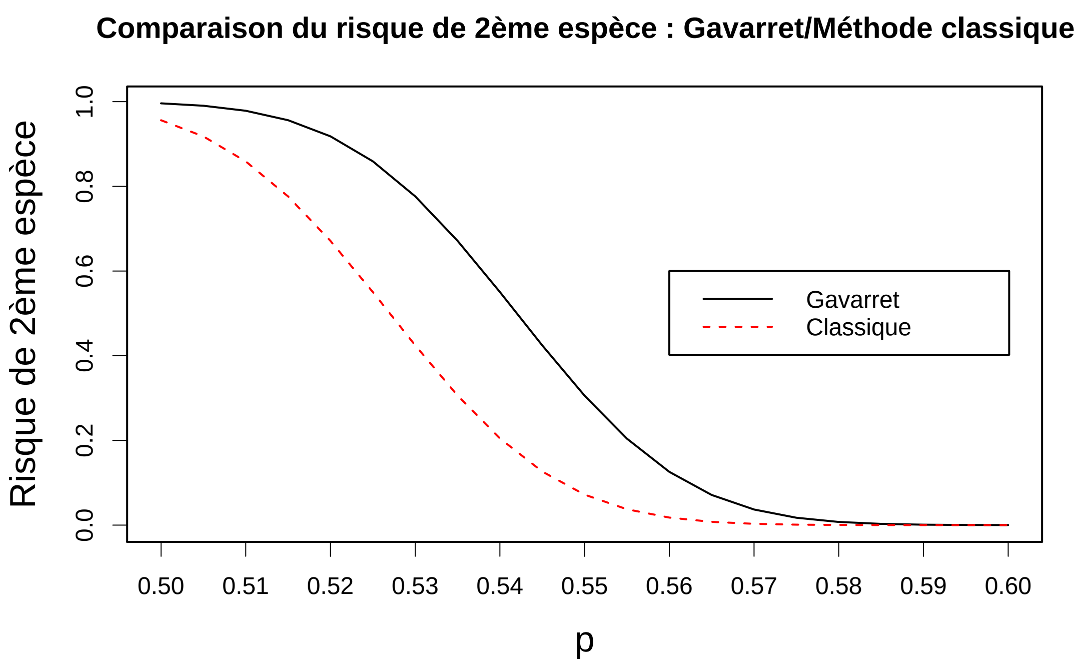
Il suffit alors de prendre pour A^* le plus petit entier x tel que : $\dfrac{(x-1) - 1000 \times 0{,}50}{\sqrt{1000 \times 0{,}50 \times 0.50}} > 1{,}65$.

On trouve : $A^* = 528$.

c) Calcul de $\beta^*(p) = \mathbb{P}_p(X < 528) = \mathbb{P}_p(X \leq 527)$.

- Pour $p = 0{,}51$, $\beta^*(p) = \mathbb{P}_p(X < 528) = 0{,}858\,900$;
- Pour $p = 0{,}52$, $\beta^*(p) = \mathbb{P}_p(X < 528) = 0{,}671\,144$;
- Pour $p = 0{,}53$, $\beta^*(p) = \mathbb{P}_p(X < 528) = 0{,}424\,623$;
- Pour $p = 0{,}54$, $\beta^*(p) = \mathbb{P}_p(X < 528) = 0{,}204\,732$;
- Pour $p = 0{,}55$, $\beta^*(p) = \mathbb{P}_p(X < 528) = 0{,}071\,874$;
- Pour $p = 0{,}60$, $\beta^*(p) = \mathbb{P}_p(X < 528) = 0{,}000\,001$.

On observe que pour $p = 0{,}51; 0{,}52; 0{,}53; 0{,}54; 0.55; 0.60$, on a $\beta^*(p) < \beta(p)$.
Puisque $(X < 528) \subset (X < 543)$, $\mathbb{P}_p(X < 528) < \mathbb{P}_p(X < 543)$ et ce résultat se généralise.



Commentaire :

1ère remarque : L'hypothèse nulle $H_0 : K \leq H$ correspond en fait à un ensemble possible de valeurs pour p qui est l'intervalle $[0\,;0{,}50]$. Pour un "risque de première espèce" donné, le "risque de deuxième espèce" est d'autant plus fort que la "vraie" valeur du paramètre p, tout en n'appartenant pas à cet ensemble, est "proche" de lui. C'est ce qu'avait bien perçu Gavarret quand il prend beaucoup de précautions pour énoncer la conclusion dans le cas où l'hypothèse nulle n'est pas rejetée (voir p. 23 de l'article et [G, p. 158]).

2ème remarque : Si on prend un risque $\alpha = 0{,}0047$, pour $p = 0{,}53$, le "risque de deuxième espèce" est égal à $\beta(0{,}53) = 0{,}776\,467$ alors que si on prend un risque $\alpha = 0{,}05 > 0{,}0047$, le "risque de deuxième espèce" est égal à $\beta^*(0{,}53) = 0{,}424\,623 < 0{,}776\,467 = \beta(0{,}53)$. De façon générale, pour une même valeur du paramètre inconnu, lorsque le "risque de première espèce" diminue, le "risque de deuxième espèce" augmente.

Il n'est pas malheureusement pas possible de "gagner sur les deux tableaux". Dans son ouvrage, Gavarret s'est limité à prendre de fait un "risque de première espèce" égal à $0{,}0047$ car il a conservé les formules de Poisson. Le "risque de seconde espèce" n'est jamais calculé mais Gavarret sent confusément que celui-ci n'est pas égal au "risque de première espèce" dans les conclusions qu'il donne à plusieurs de ses tests. Ce "risque de seconde espèce" ne pourrait l'être exactement que si on connaissait la véritable valeur de p, or celle-ci ne peut être qu'estimée.
Si ici, on l'estime par $\widehat{p} = \dfrac{900}{1000}$, il est facile de montrer que $\beta(\widehat{p}) \simeq 0$.

2.4.5. Soit $p = \dfrac{K}{K+H}$. On note x le nombre de blanches tirées.

Première interprétation : on teste $H_0 : p \leq 0{,}75$ contre $H_1 : p > 0{,}75$.

Règle de décision : on rejette $H_0 : p \leq 0{,}75$ si et seulement si $x \geq 765$ ou

Règle de décision : on rejette $H_0 : p \leq 0{,}75$ si et seulement si $P_{0{,}75}(X \geq x) < 0{,}0047$.

Ici, $x = 900$ donc on rejette $H_0 : p \leq 0{,}75$. Il y a "beaucoup plus de blanches que de noires". On évalue p par $\widehat{p} = 0{,}90$.

Le "risque de deuxième espèce" est évalué par : $\beta(0{,}90) = \mathbb{P}_{0{,}90}(X < 765) \simeq 0$

Deuxième interprétation : on teste $H_0 : p \leq 0{,}80$ contre $H_1 : p > 0{,}80$.

Règle de décision : on rejette $H_0 : p \leq 0{,}80$ si et seulement si $x \geq 814$ ou

Règle de décision : on rejette $H_0 : p \leq 0{,}80$ si et seulement si $P_{0{,}80}(X \geq x) < 0{,}0047$

Ici, $x = 900$ donc on rejette $H_0 : p \leq 0{,}80$. Il y a "beaucoup plus de blanches que de noires". On évalue p par $\widehat{p} = 0{,}90$.

Le "risque de deuxième espèce" est évalué par : $\beta(0{,}90) = \mathbb{P}_{0{,}90}(X < 814) \simeq 0$.

Troisième interprétation : on teste $H_0 : p \leq 0{,}90$ contre $H_1 : p > 0{,}90$.

Règle de décision : on rejette $H_0 : p \leq 0{,}80$ si et seulement si $x \geq 911$ ou

Règle de décision : on rejette $H_0 : p \leq 0{,}80$ si et seulement si $P_{0.90}(X \geq x) < 0{,}0047$

Ici, $x = 900$ donc on ne rejette pas $H_0 : p \leq 0{,}80$. On ne peut pas décider : “il y a beaucoup plus de blanches que de noires”. On évalue p par $\widehat{p} = 0{,}90$.

Le “risque de deuxième espèce” est évalué par : $\beta(0{,}90) = P_{0.90}(X < 911) = 0{,}854$.

Ainsi donc, puisqu'on ne rejette pas H_0, on est tenté de décider qu' “il n'y a pas beaucoup plus de blanches que de noires”. Mais dans ce cas on a $85{,}4\%$ de chances de se tromper.

> *Commentaire :* Cette activité illustre la nécessité d'être précis dans la définition de l'alternative qui se pose au statisticien.

Indications sur l'activité 2.5

2.5.1 D'après la théorie des intervalles de confiance pour une proportion inconnue :

$\mathbb{P}(\widehat{p} - \epsilon < p < \widehat{p} + \epsilon) = 1 - \alpha$ avec $\epsilon = z_\alpha \sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$

où z_α tel que $\mathbb{P}(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$. Ici pour $\alpha = 0{,}0047$ on a $z_\alpha \simeq 2\sqrt{2}$ donc :

$\epsilon = 2\sqrt{\dfrac{2 \times 0{,}90 \times (1 - 0{,}90)}{1000}} = 0{,}0268$ soit $2{,}68\%$.

L'intervalle de confiance pour p est donc $[0{,}8732\,; 0{,}9268]$, ou, en pourcentage, $[87{,}32\%\,; 92{,}68\%]$.

> *Commentaire :* Il s'agit simplement du calcul d'un intervalle de confiance avec un niveau différent de celui utilisé dans les classes de B.T.S. et de lycée.

2.5.2. Puisque l'on a $\left(|p - \widehat{p}| \leq 2\sqrt{\dfrac{2p(1-p)}{n}}\right) \subset \left(|p - \widehat{p}| \leq \sqrt{\dfrac{2}{n}}\right)$ et

$\mathbb{P}((|p - \widehat{p}| \leq 2\sqrt{\dfrac{2p(1-p)}{n}}) = 1 - 0{,}0047$, alors $\mathbb{P}(|p - \widehat{p}| \leq \sqrt{\dfrac{2}{n}}) \geq 0{,}9953$.

Il suffit de prendre n tel que $\epsilon = \sqrt{\dfrac{2}{n}}$ d'où $n \simeq \dfrac{2}{\epsilon^2}$.

> *Commentaire :* Il existe des méthodes plus fines pour optimiser la taille de l'échantillon connaissant l' “erreur” que l'on accepte entre la “réalité” et la proportion observée, mais celle-ci fournit une première estimation de cette taille. Si, avec le niveau de confiance implicite pris par Gavarret, on accepte une “erreur” de 1% autour du pourcentage observé, il faut effectuer un sondage sur 20 000 personnes !

Indications sur l'activité 2.6

2.6.1. $\binom{20}{20} = 1, \qquad \binom{20}{19} = 20, \qquad \binom{20}{18} = 190, \qquad \binom{20}{17} = 1140, \qquad \binom{20}{16} = 4845$

2.6.2.

- $\mathbb{P}_{0,50}(Y \geq 16) = 0{,}005\,91$, $\mathbb{P}_{0,50}(Y \geq 17) = 0{,}001\,29$, $\mathbb{P}_{0,50}(Y \geq 18) = 0{,}000\,20$, $\mathbb{P}_{0,50}(Y \geq 19) = 0{,}000\,02$, $\mathbb{P}_{0,50}(Y \geq 20) \simeq 0$
- $\mathbb{P}_{0,75}(Y \geq 16) = 0{,}414\,84$, $\mathbb{P}_{0,75}(Y \geq 17) = 0{,}225\,15$, $\mathbb{P}_{0,75}(Y \geq 18) = 0{,}091\,26$, $\mathbb{P}_{0,75}(Y \geq 19) = 0{,}024\,31$, $\mathbb{P}_{0,75}(Y \geq 20) = 0{,}003\,17$
- $\mathbb{P}_{0,80}(Y \geq 16) = 0{,}629\,64$ $\mathbb{P}_{0,80}(Y \geq 17) = 0{,}411\,45$, $\mathbb{P}_{0,80}(Y \geq 18) = 0{,}206\,08$, $\mathbb{P}_{0,80}(Y \geq 19) = 0{,}069\,17$, $\mathbb{P}_{0,80}(Y \geq 20) = 0{,}011\,53$
- $\mathbb{P}_{0,90}(Y \geq 16) = 0{,}956\,82$, $\mathbb{P}_{0,90}(Y \geq 17) = 0{,}867\,04$, $\mathbb{P}_{0,90}(Y \geq 18) = 0{,}676\,93$, $\mathbb{P}_{0,90}(Y \geq 19) = 0{,}391\,75$, $\mathbb{P}_{0,90}(Y \geq 20) = 0{,}121\,57$.

2.6.3.

a)

1er cas : *Règle de décision 1 :* on rejette l'hypothèse nulle (et donc on accepte l'hypothèse alternative "il y a plus de blanches que de noires") si et seulement si $y \geq 17$.

Règle de décision 2 : on rejette l'hypothèse nulle (et donc on accepte l'hypothèse alternative "il y a plus de blanches que de noires") si et seulement $\mathbb{P}_{0,50}(Y \geq y) \leq 0{,}0047$.

Puisque $y = 18$, la décision sera donc : "il y a plus de blanches que de noires".

Le "risque de deuxième espèce" pour ce test est évalué par $\mathbb{P}_{0,90}(Y < 17) = 0{,}132\,95$.

2ème cas : *Règle de décision 1 :* on rejette l'hypothèse nulle (et donc on accepte l'hypothèse alternative "il y a plus de blanches que de noires") si et seulement si $y = 20$.

Règle de décision 2 : on rejette l'hypothèse nulle (et donc on accepte l'hypothèse alternative "il y a plus de blanches que de noires") si et seulement $\mathbb{P}_{0,75}(Y \geq y) \leq 0{,}0047$.

Puisque $y = 18$, la décision sera donc : "il n'y a pas trois fois plus de blanches que de noires". Le "risque de deuxième espèce" pour ce test est évalué par $\mathbb{P}_{0,90}(Y < 20) = 0{,}878\,42$.

3ème cas : puisque $\mathbb{P}_{0,80}(Y = 20) = 0{,}011\,53 > 0{,}0047$, il n'est pas possible de construire un test ayant $\alpha = 0{,}0047$ comme "risque de première espèce".

4ème cas : puisque $\mathbb{P}_{0,90}(Y = 20) = 0{,}12157 > 0,0047$, il n'est pas possible de construire un test ayant $\alpha = 0{,}0047$ comme "risque de première espèce".

Dans tous les cas, que ce soit avec un tirage de 1000 boules ou avec un tirage de 20 boules, la proportion p est estimée par $0{,}90$.

b) Pour le premier cas, avec un tirage de 1000 boules, le "risque de deuxième espèce" est évalué par $\mathbb{P}_{0,90}(X < 543) \simeq 0$ alors qu'ici, pour le même cas, mais avec un tirage de 20 boules, ce risque est évalué à $\mathbb{P}_{0,90}(Y < 17) = 0{,}1329$.

Pour le deuxième cas, avec un tirage de 1000 boules, le "risque de deuxième espèce" est évalué par $\mathbb{P}_{0,90}(X < 765) \simeq 0$ (voir 2.4.4.) alors qu'ici, pour le même cas, mais avec un tirage de 20 boules, ce risque est évalué à $\mathbb{P}_{0,90}(Y < 20) = 0{,}8784$.

Cet exemple permet donc d'observer que le "risque de deuxième espèce" dépend de n. Plus n est grand, plus le "risque de deuxième espèce" diminue.

c) Soit le problème de test : $H_0 : K \leq H$ contre $H_1 : K > H$ avec un "risque de premier espèce" $\alpha = 0{,}0047$.

Posons $p_0 = \dfrac{K}{N}$ et soit A_n le plus petit entier y tel que $\mathbb{P}_{p_0}(Y \geq y) \leq \alpha$.

Rappelons que $F_Z(2{,}597) = 1 - 0{,}0047$. On a vu que $A_n \sim np_0 + 2{,}597\sqrt{np_0(1-p_0)} + 1$

Notons le "risque de deuxième espèce" $\beta_n(p) = \mathbb{P}_p(Y < A_n)$.

$$\begin{aligned}\beta_n(p) &= \mathbb{P}_p\left(Y \leq np_0 + 2{,}597\sqrt{np_0(1-p_0)}\right)\\ &= \mathbb{P}_p\left(\frac{Y-np}{\sqrt{np(1-p)}} \leq \frac{np_0 + 2{,}597\sqrt{np_0(1-p_0)} - np}{\sqrt{np(1-p)}}\right)\\ &= F_Z\left(\frac{\sqrt{n}(p_0-p)}{\sqrt{p(1-p)}} + 2{,}597\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}}\right)\end{aligned}$$

Pour $p > p_0$, $\dfrac{\sqrt{n}(p_0-p)}{\sqrt{p(1-p)}} + 2{,}597\dfrac{\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}}$ tend vers $-\infty$ quand n tend vers l'infini.

Donc $\beta_n(p)$ tend vers 0 quand n tend vers l'infini.

La puissance du test $l_n(p) = 1 - \beta_n(p)$, probabilité d'avoir confiance dans la décision quand celle-ci est le rejet de l'hypothèse nulle, converge donc vers 1 quand n tend vers l'infini.

2.6.4.

1er problème : Aujourd'hui, avec le risque $\alpha = 0{,}05$, puisque $x = 532 > 528$, on rejetterait l'hypothèse nulle donc, on conclurait : "il y a plus de blanches que de noires dans l'urne".

Parce que $x = 532 < 543$, avec le risque pris par Gavarret, on ne pourrait rejeter l'hypothèse nulle et on ne conclurait pas : "il y a plus de blanches que de noires dans l'urne".

En utilisant la méthode de la **p-value**, on obtiendrait la même conclusion.

En effet, ici **p-value** $= \mathbb{P}(X \geq 532) = 0{,}0249$. Puisque **p-value** $= 0{,}0249 < 0{,}05$, aujourd'hui, on rejetterait l'hypothèse nulle et, avec le risque pris par Gavarret, puisque **p-value**$=0{,}0249 > 0{,}0047$, on ne rejetterait pas l'hypothèse nulle.

Avec les conventions actuelles rappelées dans le commentaire du 2.4.3., on conclurait que le nombre de blanches dans l'urne est **significativement** plus élevé que le nombre de noires, car $0{,}01 < 0{,}0249 < 0{,}05$.

2ème problème : Soit T la variable aléatoire qui, à un tirage de 2000 boules dans une urne contenant un proportion p de blanches associe le nombre de boules blanches obtenu. $T \sim \mathcal{B}(2000, p)$.

Pour utiliser la deuxième règle de décision, il suffit de calculer $\mathbb{P}_{0{,}50}(T \geq 1064)$. On utilise l'approximation normale.

p-value $= \mathbb{P}_{0{,}50}(T \geq 1064) = 1 - \mathbb{P}_{0{,}50}(T < 1064) = 1 - \mathbb{P}_{0{,}50}(T \leq 1063)$

$$= 1 - F_Z\left(\frac{1063 - 2000 \times 0{,}50}{\sqrt{2000 \times 0{,}50 \times 0{,}50}}\right) = 1 - F_Z(2{,}817) = 0{,}0024$$

Puisque **p-value <** $0{,}05$, aujourd'hui on rejetterait l'hypothèse nulle donc, on conclurait : "il y a plus de blanches que de noires dans l'urne".

Puisque **p-value <** $0{,}0047$, avec le risque pris par Gavarret, on rejetterait l'hypothèse nulle donc on conclurait : "il y a plus de blanches que de noires dans l'urne". Avec les conventions actuelles rappelées dans le commentaire du 2.4.3., on conclurait que le nombre de blanches dans l'urne est **très significativement** plus élevé que le nombre de noires, car $0{,}001 < 0{,}0024 < 0{,}01$.

Première remarque importante :

Dans les deux cas, si on estime la valeur du paramètre inconnu p (composition de l'urne) par la proportion observée, celle-ci est la même dans les deux problèmes c'est-à-dire $\widehat{p} = 0{,}532 = 53{,}2\%$.

On constate donc que, dans les deux cas, cette valeur est aussi éloignée de l'ensemble des valeurs correspondant à l'hypothèse nulle soit $[0\,; 0{,}50]$. Or dans le premier problème, on énonce une conclusion "**significative**" alors que dans le deuxième problème cette conclusion est "**très significative**".

A l'écoute de ces conclusions différentes, il est souvent entendu que ceci impliquerait qu'il pourrait y avoir beaucoup plus de blanches dans l'urne pour le deuxième cas que dans le premier cas. Ceci n'est pas exact.

Le qualificatif de "**significatif**" ou "**très significatif**" ne concerne que le niveau de confiance portée à la conclusion qui dépend de la taille de l'échantillon.

C'est pourquoi lorsque deux laboratoires différents constatent pour l'un un effet "**significatif**" et dans l'autre un effet "**très significatif**", avant de conclure que l'effet est meilleur pour le dernier laboratoire, il faut s'assurer que les tailles des échantillons observés sont identiques. Cette remarque vaut bien sûr à fortiori quand un des laboratoires déclare une absence d'effet et l'autre un effet "**significatif**".

Deuxième remarque :

Au lieu de raisonner sur x nombre de boules blanches tirées dans un tirage de n boules, il est aussi possible de le faire sur la proportion observée $\widehat{p} = \dfrac{x}{n}$.

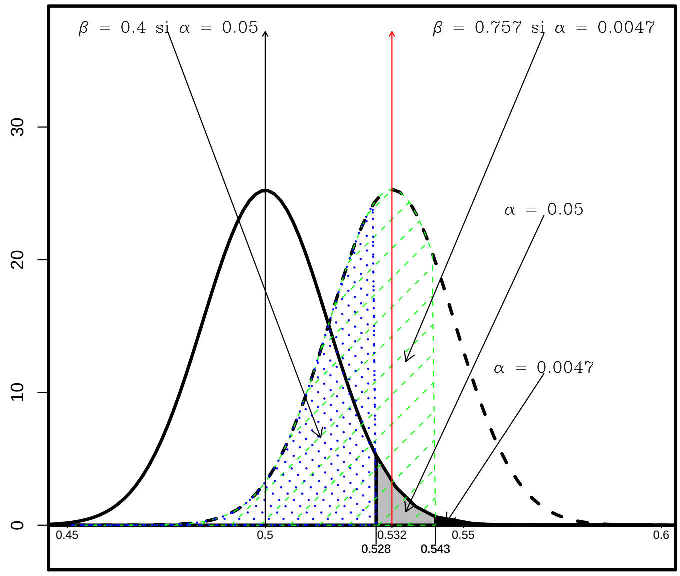
Pour connaître la proportion p^* telle que $\widehat{p} \geq p^*$ entraîne le rejet de $H_0 : p \leq 0{,}50$ avec une probabilité inférieure à α, il suffit de trouver A^* tel que $x \geq A^*$ entraîne le rejet de $H_0 : p \leq 0{,}50$ avec une probabilité inférieure à α et poser $p^* = \frac{A^*}{n}$.

Pour exemple, avec le risque $\alpha = 0{,}0047$ et un tirage de $n = 1000$ boules, on avait trouvé $A^* = 543$. Donc $p^* = 0{,}543$, et si la proportion observée est supérieure ou égale à $0{,}543$, la décision sera le rejet de $H_0 : p \leq 0{,}50$.

De même, avec le risque $\alpha = 0{,}0047$ et un tirage de $n = 2000$ boules, on trouve $A^* = 1060$. Donc $p^* = 0{,}530$, et si la proportion observée est supérieure ou égale à $0{,}530$, la décision sera le rejet de $H_0 : p \leq 0{,}50$.

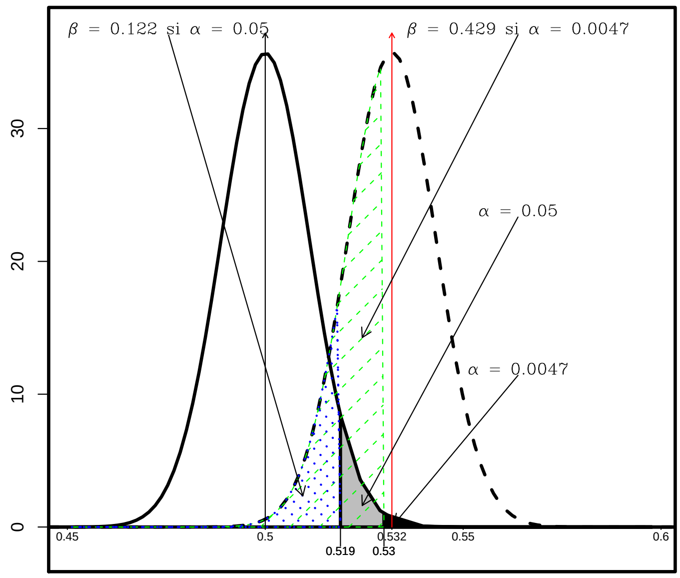
Les deux figures (à la même échelle) de la page suivante illustrent le cas où, avec des tailles d'échantillons différentes, mais avec une même proportion observée $\widehat{p} = 0{,}532$, les décisions (entre guillemets) peuvent être différentes suivant le "risque de première espèce" choisi et la taille de l'échantillon.

Test H0 : p<= 0.5 contre H1 : p> 0.5



Résultat : 532 blanches parmi 1000 boules tirées

Test H0 : p<= 0.5 contre H1 : p> 0.5



Résultat : 1064 blanches parmi 2000 boules tirées

	Décision "Gavarret"	Décision "aujourd'hui"
1000 boules	"$p \leq 0{,}50$" Risque de se tromper évalué à $0{,}757$	"$p > 0{,}50$" Risque de se tromper $\leq 0{,}05$
2000 boules	"$p > 0{,}50$" Risque de se tromper $\leq 0{,}0047$	"$p > 0{,}50$" Risque de se tromper $\leq 0{,}05$

Gavarret a bien sûr raison de ne pas se sentir éclairé par un nombre de tirages aussi faible mais son exigence d'une centaine d'observations doit être révisée.

2.6.5. : Pour montrer qu'il n'est pas équivalent d'effectuer le test $H_0 : K \leq H$ contre $H_1 : K > H$ et le test $H_0 : K \geq H$ contre $H_1 : K < H$, il faut déterminer la zone de rejet dans le cas de ce dernier test.

Pour choisir entre les deux hypothèses : $K \geq H$ et $K < H$, la règle de décision est implicitement la suivante : "Soit X la variable qui associe à un tirage de 1000 boules le nombre x de blanches. On décide $K < H$ si $x \leq A'$, où A' est un nombre tel que, si en réalité on avait $K \geq H$ la probabilité de l'événement $(X \leq A')$ serait très petite" (ici, inférieure à $0{,}0047$).

L'utilisation des tables montre que $F_Z(2{,}597) = 0{,}9953$. Puisque la symétrie de la densité de probabilité de Z, variable suivant une loi normale centrée réduite, entraîne $\mathbb{P}(Z \leq -2{,}597) = 0{,}0047$. On peut trouver A' en cherchant le plus grand entier x tel que $x - 1000 \times 0{,}50 \leq -2{,}597\sqrt{1000 \times 0{,}50 \times 0{,}50}$. Puisque $500 - 2{,}597\sqrt{1000 \times 0{,}50 \times 0{,}50} = 458{,}937$, on trouve $A' = 458$.

Pour ce nouveau problème de test, la règle de décision est donc la suivante :

On rejette $H_0 : K \geq H$ si $x \leq 458$.

Comparaison des décisions (entre guillemets) suivant le test et x nombre de boules blanches tirées dans une urne avec 1000 boules.

Rappel : Pour le test $H_0 : K \leq H$ contre $H_1 : K > H$, on rejette $H_0 : K \leq H$ si $x \geq 543$.

x	Test $H_0 : K \leq H$ contre $H_1 : K > H$	Test $H_0 : K \geq H$ contre $H_1 : K < H$
550	"$K > H$"	"$K \geq H$"
502	"$K \leq H$"	"$K \geq H$"
480	"$K \leq H$"	"$K \geq H$"
420	"$K \leq H$"	"$K < H$"

On remarque :

1. Pour un tirage de 502 boules blanches, les conclusions sont contradictoires. Il en est de même pour un tirage de 480 boules blanches.
2. Pour un tirage de 550 boules blanches, en décidant "$K > H$" avec le test $H_0 : K \leq H$ contre $H_1 : K > H$, le risque de se tromper est inférieur à $0{,}0047$ alors qu'avec le test $H_0 : K \geq H$ contre $H_1 : K < H$ le risque de se tromper est celui que l'on prend en ne rejetant pas l'hypothèse nulle (donc le risque de deuxième espèce que l'on ne peut pas contrôler).
3. Pour un tirage de 420 boules blanches, en décidant $H_0 : K \geq H$ contre $H_1 : K < H$ avec le test $H_0 : K \geq H$ contre $H_1 : K < H$, le risque de se tromper est inférieur à $0{,}0047$ alors qu'avec le test $H_0 : K \leq H$ contre $H_1 : K > H$ le risque de se tromper est celui que l'on prend en ne rejetant pas l'hypothèse nulle (donc le risque de deuxième espèce que l'on ne peut pas contrôler).

Les différences s'expliquent par l'absence de symétrie entre les deux hypothèses pour la construction de la zone de rejet. Entre deux hypothèses alternatives, il faut d'abord décider celle pour laquelle on veut contrôler le risque de se tromper en affirmant qu'elle est vraie. C'est cette hypothèse qui sera choisie comme hypothèse H_1 (et par conséquent l'autre sera choisie comme H_0).

Commentaire général sur les activités 2 :
Les conclusions tirées d'un test dépendent de plusieurs éléments dont entre autres :

- la taille de l'échantillon (voir 2.6.3.) ;
- la "distance" entre la valeur du paramètre inconnu et l'ensemble des valeurs du paramètres correspondant à l'hypothèse nulle. Plus cette distance est petite, plus le risque de se tromper en ne rejetant pas l'hypothèse nulle est grand (voir 2.4.4.) ;
- la précision de l'énoncé (voir 2.4.5.) ;
- l'hypothèse pour laquelle on veut contrôler le risque de se tromper en affirmant qu'elle est vraie (voir 2.6.5.) ;
- le risque (choisi par l'expérimentateur) de se tromper en rejetant l'hypothèse nulle (voir 2.6.4.).

Gavarret avait très bien perçu l'importance des deux premiers éléments sans les formaliser mathématiquement. Quant à la construction d'intervalles de confiance, il a, dès le début de son ouvrage, posé les questions importantes qui sont encore celles posées aujourd'hui, en particulier par la pratique des sondages.

Indications sur l'activité 3.

On pose $p_\alpha = \frac{K_\alpha}{K_\alpha + H_\alpha}$.

La probabilité de l'événement : "l'individu numéro α meurt de la pathologie" est bien sûr p_α.

On note par p la probabilité de l'événement E : "un individu tiré au hasard dans la population meurt de la pathologie". Soit T la variable aléatoire qui associe au tirage au hasard le numéro de l'individu tirée.

D'après la formule des probabilités totales,
$\mathbb{P}(E) = \mathbb{P}(T=1)\mathbb{P}_{(T=1)}(E) + \mathbb{P}(T=2)\mathbb{P}_{(T=2)}(E) + \cdots + \mathbb{P}(T=\alpha)\mathbb{P}_{(T=\alpha)}(E) +$
$\cdots + \mathbb{P}(T=N)\mathbb{P}_{(T=N)}(E)$. Mais $\mathbb{P}_{(T=\alpha)}(E) = p_\alpha$ donc :

$$p = \frac{1}{N}\sum_{\alpha=1}^{N} p_\alpha.$$

Commentaire : Outre l'utilisation de la formule des probabilités totales, l'objectif est de faire comprendre la différence entre l'examen d'un malade particulier numéro α et l'examen d'un malade pris au hasard dans une population. Dans le deuxième cas, l'événement "mort de l'individu" est le résultat de deux hasards, d'abord celui résultant du tirage au hasard de l'individu dans la population et ensuite celui résultant de la "chance" de mourir de la pathologie pour cet individu, "chance" qui peut être différente suivant le malade.

Indications sur l'activité 4.

Série initiale	En ajoutant la série 1	En ajoutant la série 2
12 morts sur 30 malades	0,200	0,500
120 morts sur 300 malades	0,384	0,413
240 morts sur 600 malades	0,392	0,406
360 morts sur 900 malades	0,394	0,404
480 morts sur 1200 malades	0,396	0,403
600 morts sur 1500 malades	0,397	0,402
720 morts sur 1800 malades	0,397	0,402
840 morts sur 2100 malades	0,398	0,401

Le résultat pour la ligne $l \geq 2$ est donné :

– pour la première colonne par : $\dfrac{120(l-1)+3}{300(l-1)+20}$

– pour la deuxième colonne par : $\dfrac{120(l-1)+8}{300(l-1)+10}$

Il est aisé de montrer que chacun de ces résultats tend vers 0,40 quand l tend vers l'infini.

Commentaire : Il s'agit simplement de ne pas se contenter d'observer le rétrécissement de l'intervalle quand l'échantillon devient grand mais de le démontrer.

Indications sur l'activité 5.

Enoncé des résultats avec un niveau de confiance de $0{,}95$ pour chacun des trois cas.

1. L'erreur est égale à $0{,}0010$ d'où :

   Gavarret aurait conclu : au lieu de la proposition précédente, nous devons réellement admettre celle-ci :

   Sur $10\ 000$ naissances, le nombre des garçons varie entre $\begin{cases} 5\ 185 \\ 5\ 165 \end{cases}$

2. L'erreur est égale à $0{,}0072$ d'où :

   Il aurait conclu : au lieu de la proposition précédente, nous devons réellement admettre celle-ci :

   Sur $10\ 000$ naissances, le nombre des garçons varie entre $\begin{cases} 5\ 140 \\ 4\ 996 \end{cases}$

3. L'erreur est égale à $0{,}0102$ d'où :

   Il aurait conclu : au lieu de la proposition précédente, nous devons réellement admettre celle-ci :

   Sur $10\ 000$ naissances, le nombre des garçons varie entre $\begin{cases} 5\ 147 \\ 4\ 943 \end{cases}$

On remarque que plus le degré de confiance diminue (on passe de $0{,}9953$ à $0{,}95$), plus la longueur de l'intervalle diminue et donc plus la précision augmente.

Avec la méthode des intervalles de confiance avec un niveau de confiance égal à $0{,}9953$, dans le premier exemple, l'hypothèse nulle aurait été rejetée donc on aurait énoncé la proposition suivante : "le nombre de garçons qui naissent est différent de celui du nombre de filles qui naissent". La probabilité de se tromper en énonçant ce résultat est inférieure à $0{,}0047$.

Avec la méthode des intervalles de confiance de niveau de confiance égal à $0,9953$, dans le deuxième et dans le troisième exemple, l'hypothèse nulle n'aurait pas été rejetée donc on aurait pu énoncer la proposition suivante : "lorsqu'un enfant naît, il y a autant de chance que ce soit un garçon qu'une fille" mais avec un "risque de deuxième espèce" qu'il convient d'estimer.

Ces conclusions restent les mêmes dans le cas d'un "risque de premier espèce" égal à $\alpha = 0{,}05$ (ou un niveau de confiance égal à $0{,}95$).

Commentaire : Ici encore, il s'agit de faire comprendre combien le niveau de confiance de l'intervalle est déterminant pour la construction de l'intervalle. Une question pourrait être posée : quel intervalle pour un niveau de confiance égal à 1 ?

Indications sur l'activité 6.

$\widehat{p_1} = 0{,}516\,431, \quad \widehat{p_2} = 0{,}517\,526$ donc $d = |\widehat{p_1} - \widehat{p_2}| = 0{,}001\,095$.

$$l = 2\sqrt{\frac{2 \times 417\,490 \times 441\,488}{912\,978^3} + \frac{2 \times 468\,151 \times 436\,443}{904\,594^3}} = 0{,}002\,036.$$

Donc $d < l$. La différence reste bien inférieure à la limite, ce qui justifie la décision de Gavarret.

Commentaire : Ici, il s'agit simplement de faire manipuler la formule de Poisson.

Indications sur l'activité 7.

7.1. Il faut bien sûr que $b > 0$ pour que f soit croissante.

Deux exemples qui ne permettent pas une telle modélisation :

Exemple 1 : $a = 0$ et $b = 0{,}2$. Pour $x > 5$, $f(x) > 1$. Or une probabilité, par définition, ne peut être strictement supérieure à 1. Donc cette modélisation ne convient pas.

Exemple 2 : $a = -0{,}5$ et $b = 0{,}1$. Pour $x < 5$, $f(x) < 0$. Or une probabilité, par définition, ne peut être strictement négative. Donc cette modélisation ne convient pas.

7.2. a) Soit f définie par : $f(x) = \dfrac{e^{a+bx}}{1+e^{a+bx}}$

Une étude de la fonction montre que $0 < f(x) < 1$ et que, si $b > 0$, cette fonction est croissante.

Soit f définie par : $f(x) = F_Z(a + bx)$ où F_Z est la fonction de répartition de la loi normale réduite centrée.

Puisque $0 < F_Z(z) < 0$ pour tout z réel, alors $0 < f(x) < 1$. F_Z est croissante donc f aussi.

Il suffit de calculer h fonction réciproque de g où $g(u) = \dfrac{e^u}{1+e^u}$. On trouve facilement h par : $h(v) = \ln\Big(\dfrac{v}{1-v}\Big)$. Or $f(x) = g(a + bx)$ donc si $y = f(x)$ alors $a + bx = \ln\Big(\dfrac{y}{1-y}\Big)$.

D'où la relation :

$$a + bx = \ln\Big(\frac{p_x}{1-p_x}\Big)$$

b) Ici on estime p_{x_1} par $\widehat{p_{x_1}}$ et p_{x_2} par $\widehat{p_{x_2}}$.

D'où 2 équations à 2 inconnues a et b :

$$\begin{cases} a + bx_1 = \ln\Big(\dfrac{\widehat{p_{x_1}}}{1-\widehat{p_{x_1}}}\Big) \\ a + bx_2 = \ln\Big(\dfrac{\widehat{p_{x_2}}}{1-\widehat{p_{x_2}}}\Big) \end{cases}$$

c) En appliquant à $x_1 = log_{10}(2) = 0{,}30103$ et $x_2 = log_{10}(3) = 0{,}477\,121$, on obtient $\widehat{p_{x_1}} = \dfrac{12}{20} = 0{,}60$ et $\widehat{p_{x_2}} = \dfrac{18}{25} = 0{,}72$. Et ensuite : $a = -0{,}516$ et $b = 3{,}061$.

D'où les relations : $p_x = \dfrac{e^{-0{,}516+3{,}061x}}{1+e^{-0{,}516+3{,}061x}}$ et $-0{,}516 + 3{,}061x = \ln\Big(\dfrac{p_x}{1-p_x}\Big)$.

En remplaçant p_x par $0{,}80$, on obtient $x = 0{,}621\,461$ d'où la dose = $10^{0{,}621\,461} = 4{,}18mg$.

d) Gavarret aurait certainement contesté ce résultat au vu du peu de malades testés pour chacune des doses.

7.3. Dans le cas d'une suite de données : $\widehat{p_{x_1}}$, $\widehat{p_{x_2}}$,...,$\widehat{p_{x_k}}$ pour k log-doses différentes x_1, x_2,...,x_k, une technique souvent utilisée est celle de du calcul de a et b à l'aide d'une régression linéaire simple, mais cette technique comporte des défauts, en particulier celui de provoquer un biais sur l'estimation de a et de b.

Commentaire : L'objectif visé par cette activité, outre l'étude de fonctions et la résolution d'un système de deux équations à deux inconnues, est surtout de montrer que le problème fondamental en médecine est de relier la probabilité d'apparition d'un phénomène à un facteur que l'on peut contrôler ou non (ici la dose administrée). Ceci nécessite une modélisation qui n'est pas unique.

Indications sur l'activité 8.

Mortalité moyenne fournie par la statistique	STATISTIQUES de 950 cas	
	Répartition des malades	Erreur possible
0,340 000	323 morts 627 guéris	0,043 471

Gavarret aurait conclu : il en résulte que la véritable mortalité cherchée est comprise entre les nombres : $0{,}340\,000$ plus $0{,}043\,471$, c'est-à-dire $0{,}383\,471$ et $0{,}340\,000$ moins $0{,}043\,471$, c'est-à-dire $0{,}296\,529$.

Commentaire : Ici, il ne s'agit que de faire comprendre le fonctionnement de la table et sa construction.

Indications sur l'activité 9.

Calcul de l avec la méthode actuelle.

- Pour la 1ère version, $l = 0{,}053\,231$. On a toujours $0{,}06 = d > l$, d'où la même conclusion que Gavarret.
- Pour la 2ème version, $l = 0{,}053\,231$. On a toujours $0{,}02 = d < l$, d'où la même conclusion que Gavarret.

On obtient la même limite car dans les deux cas, les tailles des deux échantillons restent les mêmes et la proportion commune p^* est estimée par la même quantité : $p^* = \dfrac{460}{1540} = 0{,}298\,701$.

Commentaire : La raison principale qui amène aujourd'hui à utiliser la formule $l = 2\sqrt{\left(\dfrac{1}{\mu} + \dfrac{1}{\mu'}\right)\left(\dfrac{2m^*n^*}{\mu^{*2}}\right)}$ plutôt que celle de Poisson est la suivante. On a vu dans l'activité 2 que la construction d'un test s'opérait en considérant au départ que l'hypothèse nulle était vraie. Or dans le test $H_0 : p_1 = p_2$ contre $H_1 : p_2 \neq p_2$, sous l'hypothèse nulle, $p_1 = p_2$ donc il est préférable d'utiliser cette information et plutôt que d'estimer séparement p_1 et p_2, il vaut mieux estimer la probabilité commune p^*. On peut démontrer mathématiquement que cette méthode est plus "puissante" que celle de Poisson car le "risque de deuxième espèce" est plus petit avec la nouvelle formule qu'avec la formule de Poisson.

Indications sur l'activité 10.

On calcule les intervalles de confiance donc de niveau $0{,}9953$ pour p_1 probabilité de mort avec la 1ère médication et pour p_2 probabilité de mort avec la 2ème médication, grâce à la formule de Poisson.

– Pour la 1ère version, l'intervalle de confiance pour p_1 est $[0{,}1642\ ;0{,}2358]$ et l'intervalle de confiance pour p_2 est $[0{,}2208\ ;0{,}2992]$. Ces intervalles se chevauchent, donc d'après la règle de décision, on n'aurait pas repoussé l'hypothèse nulle, **contrairement** à la décision de Gavarret.
– Pour la 2ème version, l'intervalle de confiance pour p_1 est $[0{,}1829\ ;0{,}2570]$ et l'intervalle de confiance pour p_2 est $[0{,}2018\ ;0{,}2782]$. Ces intervalles se chevauchent, donc d'après la règle de décision, on n'aurait pas repoussé l'hypothèse nulle, **conformément** à la décision de Gavarret.
– Pour la 3ème version, l'intervalle de confiance pour p_1 est $[0{,}2113\ ;0{,}2887]$ et l'intervalle de confiance pour p_2 est $[0{,}1549\ ;0{,}2251]$. Ces intervalles se chevauchent, donc d'après la règle de décision, on n'aurait pas repoussé l'hypothèse nulle, **contrairement** à la décision de Gavarret.

Commentaire : Cette activité permet de "tordre le cou" à une pratique souvent répandue et qui n'a pas de justification mathématique.

Indications sur l'activité 11.

Dans les deux situations, la “chance moyenne” est égale à $p = 0{,}80$ avant et elle est égale à $0{,}60$ après l'application de la mesure donc un gain de $0{,}20$. Dans la deuxième situation, la probabilité de contracter la maladie a baissé de $0{,}20$ pour chaque individu ce qui n'est pas le cas dans la première situation.

Cas général.

Soit $p = \dfrac{1}{N}\displaystyle\sum_{\alpha=1}^{N} p_\alpha$ la “chance moyenne” de contracter la maladie avant la mesure.

Soit $p^* = \dfrac{1}{N}\displaystyle\sum_{\alpha=1}^{N} p^*_\alpha$ la “chance moyenne” de contracter la maladie après la mesure.

Le gain en terme de chance moyenne est égal à $\dfrac{1}{N}\displaystyle\sum_{\alpha=1}^{N} e_\alpha$.

Ce gain peut cacher des disparités importantes entre individus. Or, la position du médecin qui se trouve devant l'individu α est de faire baisser la probabilité qu'il contracte la maladie. Pour savoir si une mesure est capable d'assurer un gain individuel qui soit égal au gain global sur toute la population, il est nécessaire d'introduire dans l'analyse un modèle dit “modèle mixte” où interviennent au moins deux facteurs de variabilité : variabilité des probabilités entre individus et, pour un individu donné, variabilité du résultat “a contracté la maladie” suivant une loi déterminée par la probabilité affectée à l'individu.

Si pour chaque individu $e_\alpha = e$, alors le gain global sera égal à e.

Le modèle $e_\alpha = e$ est mis en défaut s'il existe α tel que $p_\alpha - e < 0$. Puisque pour cet individu α, on a $p^*_\alpha = p_\alpha - e$ et donc $p^*_\alpha < 0$. Or une probabilité ne peut être négative !

> *Commentaire :* Cette activité peut être une des explications au peu d'intérêt accordé à l'époque par les médecins aux travaux de Gavarret. En effet, un bénéfice en terme de santé publique ne se répercute pas automatiquement de la même façon sur chacun des individus concernés. Gavarret avait prévenu l'objection en disant que la statistique ne pouvait prédire un événement pour un individu particulier.

Indications sur l'activité 12.

En effet, si $0{,}5 - \sqrt{\dfrac{2}{\mu}} \leq \dfrac{n_1}{n_1+n_2} \leq 0{,}5 + \sqrt{\dfrac{2}{\mu}}$, alors $1 - (0{,}5 - \sqrt{\dfrac{2}{\mu}}) \geq 1 - \dfrac{n_1}{n_1+n_2} \geq 1 - (0{,}5 + \sqrt{\dfrac{2}{\mu}})$ d'où $0{,}5 - \sqrt{\dfrac{2}{\mu}} \leq \dfrac{n_2}{n_1+n_2} \leq 0{,}5 + \sqrt{\dfrac{2}{\mu}}$.

> *Commentaire :* Cette “indifférence” ne vaut ici que dans le cas où on veut tester $H_0 : p = 0{,}50$ contre $H_1 : p \neq 0{,}50$ mais dans des cas où la valeur de référence n'est plus $0{,}50$, des précautions doivent être prises.

Indications sur l'activité 13.

Soit Y la variable aléatoire qui associe à un individu, le jour de la semaine où il est rentré à l'hôpital.

Soit $J_1 \subset \{1, 2, 3, 4, 5, 6, 7\}$, un sous-ensemble de k_1 jours et $J_2 \subset \{1, 2, 3, 4, 5, 6, 7\}$ un sous-ensemble de $k_2 = 7 - k_1$ jours avec J_1 et J_2 disjoints.

Si le résultat est uniquement dû au hasard, alors $\mathbb{P}(Y \in J_1) = \dfrac{k_1}{7}$. Si le résultat n'est pas uniquement dû au hasard, alors $\mathbb{P}(Y \in J_1) \neq \dfrac{k_1}{7}$. Si $J = J_1 \cup J_2 = \{1, 2, 3, 4, 5, 6, 7\}$ numérote les jours de la semaine, alors en prenant $J_1 = \{1, 2, 4, 5\}$ et en posant $p = \mathbb{P}(Y \in J_1)$, le problème de test est donc :

$$H_0 : p = 4/7 \text{ contre } H_1 : p \neq 4/7$$

On peut utiliser le test énoncé par Gavarret [G, p. 289-290] (voir article p. 28-29).

Le test utilisant l'artifice de se ramener au même nombre de jours n'est pas équivalent à celui qui devrait être utilisé car

$$0{,}50 - \sqrt{\frac{2}{\frac{3n_1}{4} + n_2}} \leq \frac{\frac{3n_1}{4}}{\frac{3n_1}{4} + n_2} \leq 0{,}50 + \sqrt{\frac{2}{\frac{3n_1}{4} + n_2}}$$

n'est pas équivalent à

$$\frac{4}{7} - 2\sqrt{\frac{2 \times \frac{4}{7}\frac{3}{7}}{n_1 + n_2}} \leq \frac{n_1}{n_1 + n_2} \leq \frac{4}{7} + 2\sqrt{\frac{2 \times \frac{4}{7}\frac{3}{7}}{n_1 + n_2}}$$

La démonstration n'est pas immédiate. On pose $n = n_1 + n_2$ et $x = n_1$. L'artifice de Gavarret consiste, dans la formule de Poisson, à remplacer n_1 par $\frac{3}{4}n_1 = \frac{3}{4}x$ et n par $n' = \frac{3}{4}n_1 + n_2 = \frac{3}{4}x + (n - x) = n - \frac{x}{4}$.

Problème posé :

Existe-t-il $n \in \mathbb{N}$ et $x \in \mathbb{N}$ avec $x \leq n$ tels que les deux propositions suivantes ne soient pas simultanément vraies ?

Proposition 1 :

$$0{,}50 - \sqrt{\frac{2}{n - \frac{x}{4}}} \leq \frac{\frac{3x}{4}}{n - \frac{x}{4}} \leq 0{,}50 + \sqrt{\frac{2}{n - \frac{x}{4}}} \tag{1}$$

Proposition 2 :

$$\frac{4}{7} - 2\sqrt{\frac{2 \times \frac{4}{7}\frac{3}{7}}{n}} \leq \frac{x}{n} \leq \frac{4}{7} + 2\sqrt{\frac{2 \times \frac{4}{7}\frac{3}{7}}{n}} \tag{2}$$

Résolution du problème :

Dans la suite de la démonstration, nous poserons $c = (\frac{4}{7})^2 = \frac{16}{49}$.

Remarquons d'abord que, en multipliant les termes des inégalités de la proposition 2 par n et en posant :

$$I_2^{\pm} = \frac{4n}{7} \pm \sqrt{6cn},$$

la proposition 2 est équivalente à :

$$x \in I_2 = [I_2^- \,; I_2^+] \tag{3}$$

Démontrons d'abord que la proposition 1 est équivalente à dire que $x \in I_1 = [I_1^- \,; I_1^+]$, où I_1^- et I_1^+ seront des réels à déterminer.

(1) est équivalent à :

$$\frac{1}{2}-\sqrt{\frac{8}{4n-x}} \leq \frac{3x}{4n-x} \leq \frac{1}{2}+\sqrt{\frac{8}{4n-x}} \tag{4}$$

ou encore à :

$$\left|\frac{1}{2}-\frac{3x}{4n-x}\right| \leq \sqrt{\frac{8}{4n-x}} \tag{5}$$

En élevant au carré :

$$\left(\frac{1}{2}-\frac{3x}{4n-x}\right)^2 \leq \frac{8}{4n-x} \tag{6}$$

En réduisant au même dénominateur, on obtient l'inéquation :

$$49x^2+32x-56nx-128n+16n^2 \leq 0 \tag{7}$$

Le calcul du discriminant réduit $\Delta^{'}$ donne :

$\Delta^{'} = \sqrt{(16-28n)^2-49(16n^2-128n)} = \sqrt{5356n+256}$ d'où les 2 racines du polynôme en x :

$\alpha_1 = \frac{-(16-28n)-\sqrt{5356n+256}}{49}$ et $\alpha_2 = \frac{-(16-28n)-\sqrt{5356n+256}}{49}$

En simplifiant et en utilisant $c = \frac{16}{49}$, l'inéquation (6) est équivalente à :

$$\frac{4n}{7}-c-c\sqrt{1+21n} \leq x \leq \frac{4n}{7}-c+c\sqrt{1+21n} \tag{8}$$

En posant :

$I_1^{\pm} = \dfrac{4n}{7}-c \pm c\sqrt{1+21n}$,

on prouve donc que la proposition 1 est équivalente à

$$x \in I_1 = [I_1^- \,; I_1^+] \tag{9}$$

Pour démontrer qu'il existe $n \in \mathbb{N}$ et $x \in \mathbb{N}$ $(x \leq n)$ ne vérifiant pas simultanément les deux propositions, il suffit de trouver $n \in \mathbb{N}$ et $x \in \mathbb{N}$ $(x \leq n)$ tels que $x \in I_1$ et $x \notin I_2$ soit vrai, ou bien $x \notin I_1$ et $x \in I_2$ soit vrai.

Remarque 1 : Soient a et b deux nombres réels avec $a < b$. Si $b-a > 1$, il existe $x \in \mathbb{N}$ tel que $a \leq x < b$.

Il suffit de prendre $x = \lfloor b \rfloor$ où $\lfloor b \rfloor$ désigne la partie entière de b.

Remarque 2 : $(\frac{4n}{7}-\sqrt{6cn})-(\frac{4n}{7}-c-c\sqrt{1+21n}) > 0$.

Montrons d'abord $c\sqrt{1+21n}-\sqrt{6cn} > 0$. En effet, $\dfrac{c\sqrt{21}}{\sqrt{6c}} = \sqrt{\dfrac{3\times7\times4\times4}{2\times3\times7\times7}} = \sqrt{\dfrac{8}{7}} > 1$.

On en déduit : $c\sqrt{21} > \sqrt{6c}$ et $c\sqrt{1+21n}-\sqrt{6cn} > c\sqrt{21n}-\sqrt{6cn} = \sqrt{n}(c\sqrt{21}-\sqrt{6c}) > 0$.

Il suffit de remarquer que cette dernière inégalité entraîne :

$(\frac{4n}{7}-\sqrt{6cn})-(\frac{4n}{7}-c-c\sqrt{1+21n}) = c+(c\sqrt{1+21n}-\sqrt{6cn}) > c > 0$.

En prenant $b = I_2^- = \dfrac{4n}{7}-\sqrt{6cn}$ et $a = I_1^- = \dfrac{4n}{7}-c-c\sqrt{1+21n}$, la remarque 1 amène à chercher $n \in \mathbb{N}$ tel que :

$(\dfrac{4n}{7}-\sqrt{6cn})-(\dfrac{4n}{7}-c-c\sqrt{1+21n}) > 1$ ou encore tel que :

$$c+c\sqrt{1+21n}-\sqrt{6cn}-1 > 0. \tag{10}$$

Mais, puisque $c+c\sqrt{1+21n}-\sqrt{6cn}-1 > c+c\sqrt{21n}-\sqrt{6cn}-1$, il suffira de trouver $n \in \mathbb{N}$ tel que :

$$c+c\sqrt{21n}-\sqrt{6cn}-1 > 0.$$

En prenant :
$\sqrt{n} > \dfrac{1-c}{c\sqrt{21}-\sqrt{6c}}$ c'est-à-dire $n > 48{,}56199$ ou encore $n \geq 49$, on obtient le résultat recherché.

Pour $n = 49$, $I_1 = [17{,}19\,;38{,}15]$ et $I_2 = [18{,}20\,;37{,}79]$. Il suffit de prendre $x = 18$ car alors $x \in I_1$ mais $x \notin I_2$. Dans ce cas, ceci revient donc à prendre $n_1 = 18$ et $n_2 = n - x = 31$.

La valeur de la fonction de test utilisée par Gavarret, pour ces valeurs de n_1 et de n_2, est égale à : $\varphi = \dfrac{\frac{3n_1}{4}}{\frac{3n_1}{4}+n_2} = 0{,}3033$. Puisque $\varphi \in [0{,}50 - \sqrt{\dfrac{2}{\frac{3n_1}{4}+n_2}}\,;0{,}50 + \sqrt{\dfrac{2}{\frac{3n_1}{4}+n_2}}] = [0{,}288\,;0{,}712]$, l'hypothèse nulle est acceptée avec ce test.

La valeur de la fonction de test dans le cas du test adéquat, pour ces valeurs de n_1 et de n_2, est égale à : $\varphi = \dfrac{n_1}{n_1+n_2} = 0{,}367$. Puisque $\varphi \notin [\frac{4}{7} - 2\sqrt{\dfrac{2 \times \frac{4}{7}\frac{3}{7}}{n}}\,; \frac{4}{7} + 2\sqrt{\dfrac{2 \times \frac{4}{7}\frac{3}{7}}{n}}] = [0{,}371\,;0{,}771]$, l'hypothèse nulle est rejetée avec ce test.

Commentaire : L'artifice utilisé par Gavarret est encore, sous d'autres formes, en vigueur de nos jours, en particulier pour augmenter artificiellement le nombre de données. Dans beaucoup de cas, il ne peut se justifier mathématiquement.

Indications sur l'activité 14.

En utilisant l'artifice, la valeur de la statistique de test est égale à $0{,}50\,977$ et se trouve bien dans l'intervalle qui est $[0,48\,697\,;0{,}51\,303]$.

En utilisant la "bonne méthode", la valeur de la statistique de test est $0{,}58\,096$ et se trouve encore dans l'intervalle qui est $[0{,}55\,951\,;0{,}58\,335]$.

Commentaire : Il s'agit simplement de mettre en oeuvre l'artifice de Gavarret et de comparer les résultats avec une méthode plus rigoureuse.

Indications sur l'activité 15.

- Test $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.
  Ici $d = 0{,}10 < 0{,}1470 = l$. On ne rejette pas H_0. Décision : $p_1 = p_2$.
- Test $H_0 : p_2 = p_3$ contre $H_1 : p_2 \neq p_3$.
  Ici $d = 0{,}03 < 0{,}1474 = l$. On ne rejette pas H_0. Décision : $p_2 = p_3$.
- Test $H_0 : p_1 = p_3$ contre $H_1 : p_1 \neq p_3$.
  Ici $d = 0{,}07 > 0{,}0614 = l$. On rejette H_0. Décision : $p_1 \neq p_3$.

  En toute logique $p_1 = p_2$ et $p_2 = p_3$ devrait impliquer $p_1 = p_3$. Mais ici, $p_1 = p_2$ et $p_2 = p_3$ ne sont que des décisions prises après une expérience aléatoire et donc assorties d'une probabilité d'être vraie. On a vu que dans le cas d'une décision du type $p_1 = p_2$, la probabilité de se tromper pouvait être importante en particulier si la différence entre p_1 et p_2 est "petite". Les règles de la logique ne s'appliquent pas à des décisions qui ne sont pas "certaines".

Commentaire : Aujourd'hui, ce sont des méthodes du type analyse de la variance ou méthode du chi-deux qui s'appliquent pour tester : $H_0 : p_1 = p_2 = p_3$ contre H_1 : il existe au moins deux p_i différents. Cependant, lorsque l'hypothèse nulle est rejetée, se pose ensuite le problème de la comparaison des p_i et, dans l'état actuel des connaissances statistiques, il n'existe pas de test idéal. La tâche que s'était donné Gavarret de classer plusieurs "médications" n'est pas achevée !

Indications sur l'activité 16.

16.1. D'après la formule de Bayes : $\mathbb{P}_{S_1}(D) = \dfrac{\mathbb{P}(D)\mathbb{P}_D(S_1)}{\mathbb{P}(S_1)}$ et $\mathbb{P}_{S_2}(D) = \dfrac{\mathbb{P}(D)\mathbb{P}_D(S_2)}{\mathbb{P}(S_2)}$.

Si $\mathbb{P}(S_1) = \mathbb{P}(S_2) = p^*$, alors $p_1^* = \dfrac{p \times p_1}{p^*}$ et $p_2^* = \dfrac{p \times p_2}{p^*}$.

Il s'ensuit que $p_1 = p_2$ entraine $p_1^* = p_2^*$ et réciproquement.

16.2. Pour traiter les données de l'érysipèle, on peut poser les événements suivants :
- S_1 : "être un homme" ;
- S_2 : "être une femme" ;
- E : "être atteint de d'érysipèle".

On note :
- $p_1 = \mathbb{P}_E(S_1)$ probabilité d'être un homme sachant qu'on est atteint d'érysipèle ;
- $p_2 = \mathbb{P}_E(S_2)$ probabilité d'être une femme sachant qu'on est atteint d'érysipèle ;
- $p_1^* = \mathbb{P}_{S_1}(E)$ probabilité d'être atteint d'érysipèle sachant que l'on est un homme ;
- $p_2^* = \mathbb{P}_{S_2}(E)$ probabilité d'être atteint d'érysipèle sachant que l'on est une femme ;
- $p = \mathbb{P}(E)$ probabilité d'être atteint d'érysipèle.

Il s'agit de tester :
$H_0 : p_1^* = p_2^*$ contre $H_1 : p_1^* \neq p_2^*$.
On peut supposer que $\mathbb{P}(S_1) = \mathbb{P}(S_2)$ et donc tester suivant la méthode de Gavarret :
$H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$
La valeur de la statistique de test est $\dfrac{n_1}{n_1 + n_2} = 0{,}477\,0115$. Elle se trouve dans l'intervalle $[0{,}446\,394 ; 0.553\,605$
On ne peut donc pas affirmer que le genre a une influence sur le fait d'attraper ou non l'érysipèle.

Commentaire : Cette activité permet l'utilisation de la formule de Bayes. Elle amène à mieux formaliser les hypothèses en présence.