Activité 1.

20 personnes (identifiées par leur numéro $\alpha = 1, ..., 20$) notent leur tension supérieure $\xi_{1,\alpha}$ en l'absence de traitement avec un tensiomètre individuel. Les nombres ne sont connus que d'eux-mêmes. Les résultats sont les suivants (données fictives) :

Tableau 1

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

Un médicament administré à tous est censé faire baisser la moyenne du groupe de ces 20 personnes. Cette moyenne **avant** traitement, notée μ_1 n'est pas connue du médecin.

Les 20 personnes notent leur tension $\xi_{2,\alpha}$ **après** le traitement avec un tensiomètre individuel. Les nombres ne sont connus que d'eux-mêmes. La moyenne **après** traitement notée μ_2 n'est pas connue du médecin.

Les résultats sont les suivants (données fictives) :

Tableau 2

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

Le médecin se contente d'interroger $n = 5$ personnes au hasard **avant** le traitement.

Il note $x_{1,i}$ ($i = 1, ..., n$) la tension correspondant à la i-ème personne tirée et $\overline{x_1}$ la moyenne de la tension de ces $n = 5$ personnes.

Il interroge ensuite $n = 5$ personnes au hasard **après** le traitement. Il note $x_{2,i}$ ($i = 1, ..., n$) la tension correspondant à la i-ème personne tirée et $\overline{x_2}$ la moyenne de la tension de ces $n = 5$ personnes.

1.1. Tirer au hasard 5 nombres compris entre 1 et 20. Attribuer à chacun de ces nombres la tension correspondant au tableau 1 et calculer la moyenne observée $\overline{x_1}$. Tirer de nouveau au hasard 5 nombres compris entre 1 et 20. Attribuer à chacun de ces nombres la tension correspondant au tableau 2 et calculer la moyenne observée $\overline{x_2}$. Au vu de ces deux séries de 5 résultats, qu'auraient décidé les tenants de la "méthode numérique" ?

1.2. Calculer maintenant μ_1 et μ_2. Construire un tableau de 20 nombres $\epsilon_{1,\alpha}$ ($\alpha = 1, ..., 20$) qui sont les écarts des tensions **avant** traitement à la moyenne μ_1. Construire de même un tableau de 20 nombres $\epsilon_{2,\alpha}$ $\alpha = 1, ..., 20$) qui sont les écarts des tensions **après** traitement à la moyenne μ_2. Les "erreurs" se compensent-elles ? Pouvait-on prévoir le résultat ? Que penser alors de la décision des tenants de la "méthode numérique" ?

1.3. Pour les 5 personnes tirées au hasard **avant** traitement,

$$\text{calculer } e_{1,i} = x_{1,i} - \mu_1 \ (i = 1, ..., 5) \text{ et } \sum_{i=1}^{5} e_{1,i}.$$

Pour les 5 personnes tirées au hasard **après** traitement,

$$\text{calculer } e_{2,i} = x_{2,i} - \mu_2 \ (i = 1, ..., 5) \text{ et } \sum_{i=1}^{5} e_{2,i}.$$

Les "erreurs" se compensent-elles ? Si ce n'est pas le cas, pouvez-vous trouver cependant des tirages de 5 individus avant le traitement et des tirages de 5 individus après le traitement pour lesquels les "erreurs" se compensent.

Gavarret avait-il raison en émettant des doutes sur le fait que les "erreurs" puissent se compenser ?

Dans le cas où "les erreurs" se compensent, pourquoi peut-on dire que $\overline{x_1} > \overline{x_2}$ (réciproquement $\overline{x_1} < \overline{x_2}$) entraîne $\mu_1 > \mu_2$ (réciproquement $\mu_1 < \mu_2$) ? Pour quelle raison cette implication ne peut elle pas se produire dans tous les cas ?

1.4. Dans le cas où vous auriez été dans la situation du médecin qui ne peut obtenir que 5 données avant traitement et 5 données après traitement, auriez-vous procédé comme lui ? Pourquoi ?

Activités 2. - Sommaire

On trouve dans les pages suivantes les énoncés de six activités qui reprennent des questions soulevées par Gavarret dans les deux premiers articles du chapitre premier de son ouvrage.
L'objet de ce chapitre est de mettre en évidence la « Nécessité de recourir au calcul des probabilités pour suppléer l'insuffisance des règles de la logique » comme l'indique le sous-titre.
Un exemple de questions soulevées par Gavarret est illustré par l'extrait de texte suivant [G., p. 46-47].
On peut accéder aux énoncés de ces activités ou revenir à l'article grâce aux liens placés sous l'extrait, dans le bas de cette page.

1° Une urne contient un nombre inconnu de boules blanches et de boules noires ; on fait mille tirages successifs d'une boule chacun. Sur ces mille tirages, faits complètement au hasard, il sort neuf cents boules blanches et cent boules noires. A la suite d'un semblable résultat, personne n'hésiterait à affirmer que nécessairement les blanches étaient beaucoup plus nombreuses dans l'urne que les noires ; personne n'hésiterait à admettre que si toutes les boules sorties étaient remises dans l'urne et mêlées à celles qui y sont restées, dans mille nouveaux tirages, faits au hasard, on obtiendrait beaucoup plus de blanches que de noires. De semblables conclusions viendraient spontanément à l'esprit de tout le monde, même de ceux qui nient l'influence des grands nombres de faits sur nos décisions. Mais là s'arrêtent les indications fournies par les règles de la *logique*. Supposons, en effet, qu'on veuille aller plus loin, et qu'on se dise : Le résultat statistique montre que sur dix boules sorties, il y a neuf blanches et une noire, dans *quelles limites d'erreur* le rapport de blanches au nombre total des boules contenues dans l'urne, peut-il différer de l'expression neuf dixièmes fournie par l'expérience? En n'appelant à son secours que les règles de la logique, qui se croirait assez éclairé pour répondre à une semblable question? Voilà une difficulté bien faite pour arrêter l'esprit le plus juste et le plus clairvoyant.

2° Une urne contient un nombre inconnu de boules blanches et de boules noires ; on fait complètement au hasard vingt tirages successifs d'une boule chacun ; sur les vingt boules sorties, dix-huit sont blanches et deux seulement noires. Ici, comme dans l'exemple précédent, sur dix boules sorties on a amené neuf blanches et une noire. La seule différence, et elle est grave, n'existe que dans le nombre des tirages effectués. Conclura-t-on de ce résultat que certainement l'urne contenait avant les tirages beaucoup plus de blanches que de noires?

Activité 2.1 Activités 2.2

Activité 2.3 Activités 2.4

Activité 2.5 Activités 2.6

Activité 2.1

Cette activité reprend les questions posées par Gavarret pages 46 et 47.

> **1° Une urne contient un nombre inconnu de boules blanches et de boules noires ; on fait mille tirages successifs d'une boule chacun. Sur ces mille tirages, faits complètement au hasard, il sort neuf cents boules blanches et cent boules noires. A la suite d'un semblable résultat, personne n'hésiterait à affirmer que nécessairement les blanches étaient beaucoup plus nombreuses dans l'urne que les noires ; personne n'hésiterait à admettre que si toutes les boules sorties étaient remises dans l'urne et mêlées à celles qui y sont restées, dans mille nouveaux tirages, faits au hasard, on obtiendrait beaucoup plus de blanches que de noires. De semblables conclusions viendraient spontanément à l'esprit de tout le monde, même de ceux qui nient l'influence des grands nombres de faits sur nos décisions. Mais là s'arrêtent les indications fournies par les règles de la *logique*. Supposons, en effet, qu'on veuille aller plus loin, et qu'on se dise : Le résultat statistique montre que sur dix boules sorties, il y a neuf blanches et une noire, dans *quelles limites d'erreur* le rapport de blanches au nombre total des boules contenues dans l'urne, peut-il différer de l'expression neuf dixièmes fournie par l'expérience? En n'appelant à son secours que les règles de la logique, qui se croirait assez éclairé pour répondre à une semblable question? Voilà une difficulté bien faite pour arrêter l'esprit le plus juste et le plus clairvoyant.**
>
> **2° Une urne contient un nombre inconnu de boules blanches et de boules noires ; on fait complètement au hasard vingt tirages successifs d'une boule chacun ; sur les vingt boules sorties, dix-huit sont blanches et deux seulement noires. Ici, comme dans l'exemple précédent, sur dix boules sorties on a amené neuf blanches et une noire. La seule différence, et elle est grave, n'existe que dans le nombre des tirages effectués. Conclura-t-on de ce résultat que certainement l'urne contenait avant les tirages beaucoup plus de blanches que de noires?**

2.1.1. Donner un exemple de composition d'urne comprenant N boules dont K blanches et H noires telle que le résultat d'un tirage de $n_1 = 1000$ boules au hasard puisse donner $k_1 = 900$ blanches et $h_1 = 100$ noires, **mais** avec cependant $K < H$ (donner des valeurs numériques pour K, N et H).

2.1.2. Même question avec $K = H$.

Activité 2.2

2.2.1. Gavarret, à partir de l'observation de $k_1 = 900$ blanches et $h_1 = 100$ noires, en a déduit que $K > H$. Aurait-il déduit le même résultat s'il avait observé $k_1 = 901$ blanches et $h_1 = 99$ noires ? Même question avec $k_1 = 902$ blanches et $h_1 = 98$ noires ? Ne peut-on pas en déduire que c'est l'observation d' "**au moins** 900 **blanches**" qui amène à la décision : $K > H$?

2.2.2. Dans ce cas, à partir des N, K et H proposés en **2.1.**, calculer la probabilité que le tirage de $n_1 = 1000$ boules au hasard donne au moins 900 blanches.

Activité 2.3

Le résultat d'un premier tirage de n_1 boules a fourni Gavarret une information sur la composition de l'urne. Une des utilisations possibles de cette information est la suivante : la véritable proportion de boules blanches est celle fournie par le premier tirage c'est-à -dire 90%.

a) En se servant de cette interprétation, calculer la probabilité que suite à un nouveau tirage de $n_2 = 1000$ boules dans cette urne, il y ait plus de blanches que de noires.

Que pensez-vous de la phrase de Gavarret : « A la suite d'un semblable résultat, personne n'hésiterait à affirmer que nécessairement les blanches étaient beaucoup plus nombreuses dans l'urne que les noires ; personne n'hésiterait à admettre que si toutes les boules sorties étaient remises dans l'urne et mêlées à celles qui y sont restées on obtiendrait beaucoup plus de blanches que de noires » ?

b) Quelle serait la probabilité de tirer au moins 950 blanches sur les 1000 boules tirées ? Comment expliquer ce paradoxe ?

Activité 2.4

Dans la suite de son ouvrage, on verra que Gavarret considère de fait un événement comme « absolument certain » [G, p. 257] si sa probabilité est supérieure ou égale à $0{,}9953$.

Dans la suite, on dira qu'un événement est "impossible au sens de Gavarret" si sa probabilité est inférieure ou égale à $0{,}0047$.

2.4.1. a) Si dans l'urne, il y a autant de blanches que de noires, c'est-à -dire $K = H$, montrer que dans un tirage de 1000 boules, le fait de tirer au moins 900 blanches est "impossible au sens de Gavarret".

Si dans l'urne il y a autant de blanches que de noires, calculer :

- la probabilité qu'il y ait au moins 546 blanches.
- la probabilité qu'il y ait au moins 543 blanches.
- la probabilité qu'il y ait au moins 542 blanches.
- la probabilité qu'il y ait au moins 541 blanches.

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

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

Si la proportion de blanches dans l'urne est 48%, calculer la probabilité de tirer au moins 543 blanches.

On pose $\mathbb{P}_p(X \geq 543)$ la probabilité d'avoir au moins 543 boules blanches parmi 1000 tirées dans une urne contenant une proportion $p = \dfrac{K}{N}$ de blanches.

A la suite des résultats précédents, on conjecture que la probabilité de tirer au moins 543 blanches parmi 1000 est une fonction croissante de la proportion p de blanches dans l'urne.

b) Montrer rigoureusement cette conjecture.

c) En déduire que si c'est l'hypothèse $K < H$ qui est vraie, on a $\mathbb{P}_p(X \geq 543) < 0{,}0047$.

Traduction : si $X \geq 543$ est réalisé, en décidant que $K > P$ alors que c'est l'hypothèse $K \leq H$ qui est vraie, on prend donc un risque de se tromper inférieur à $\alpha = 0{,}0047$.

2.4.3. Soit maintenant le problème suivant :

On doit décider entre l'hypothèse $H_0 : K \leq H$ et l'hypothèse $H_1 : K > H$.

On prend un risque au plus égal à $\alpha = 0{,}0047$ de se tromper en décidant $K > H$ alors que c'est l'hypothèse $K \leq H$ qui est vraie. $\alpha = 0{,}0047$ est dénommé "risque de première espèce".

On effectue $n = 1000$ tirages dans l'urne et on observe x blanches.

D'après les résultats précédents, on obtient la règle de décision suivante :

Règle de décision 1 : on rejette l'hypothèse nulle (et donc on accepte l'hypothèse alternative en décidant "$K > H$") si et seulement si $x \geq 543$.

Soit la règle de décision suivante :

Règle de décision 2 : on rejette l'hypothèse nulle (et donc on accepte l'hypothèse alternative en décidant "$K > H$") si et seulement $\mathbb{P}_{0{,}50}(X \geq x) \leq 0{,}0047$.

Montrer que les deux règles de décision sont équivalentes.

2.4.4. On note toujours $p = \dfrac{K}{N}$ et $\mathbb{P}_p((X < 543)$ la probabilité de ne pas rejeter l'hypothèse nulle $K \leq H$ alors qu'en réalité on a $K > H$. C'est le risque de se tromper en décidant $K \leq H$ appelé "risque de deuxième espèce". Ce risque dépend de p ($0{,}50 < p < 1$) et est noté $\beta(p)$.

a) Calculer $\beta(p)$ pour $p = 0{,}51; 0{,}52; 0{,}53; 0{,}54; 0{,}55; 0{,}60$.

Montrer rigoureusement que, plus il y a de boules blanches dans l'urne, plus le risque de se tromper en disant "il y a moins de boules blanches que de boules noires" diminue.

Ici, le "risque de première espèce" est égal à $\alpha = 0{,}0047$. Actuellement, le "risque de première espèce" est souvent pris égal à $\alpha = 0{,}05$ (5%).

Soit maintenant le nouveau problème de test.

On doit décider entre l'hypothèse $H_0 : K \leq H$ et l'hypothèse $H_1 : K > H$.

On prend un risque au plus égal à $\alpha = 0{,}05$ de se tromper en décidant $H_1 : K > H$ alors que c'est l'hypothèse $H_0 : K \leq H$ qui est vraie.

On effectue $n = 1000$ tirages dans l'urne et on observe x blanches.

Soit A^* le plus petit entier $x \leq n$ tel que $\mathbb{P}(X \geq x) \leq 0{,}05$.

On utilise une des deux règles de décision :

Règle de décision 1 : on rejette l'hypothèse nulle $H_0 : K \leq H$ (et donc on accepte l'hypothèse alternative H_1 en décidant "$K > H$") si et seulement si $x \geq A$.

Règle de décision 2 : on rejette l'hypothèse nulle $H_0 : K \leq H$ (et donc on accepte l'hypothèse alternative H_1 en décidant "$K > H$") si et seulement $\mathbb{P}_{0,50}(X \geq x) \leq \alpha$.

b) Calculer A^* dans le cas où on prend $\alpha = 0{,}05$.

Avec cette règle de décision, lorsque $x < A^*$, on ne rejette pas $K \leq H$ puisqu'on ne prend la décision " $K > H$ est vrai" que lorsque $x \geq A^*$. On note $\beta^*(p) = \mathbb{P}_p((X < A^*)$ la probabilité de ne pas rejeter l'hypothèse nulle avec un "risque de première espèce" $\alpha = 0{,}05$.

c) Calculer $\beta^*(p)$ pour $p = 0{,}51; 0{,}52; 0{,}53; 0{,}54; 0{,}55; 0{,}60$. Pour ces valeurs de p, comparer les résultats de fonctions β et β^*. Peut-on généraliser ce résultat à tout p tel que $0{,}50 \leq p \leq 1$?

Qu'en concluez-vous concernant la comparaison des risques dans les deux cas suivants : celui où le risque de première espèce est $\alpha = 0{,}0047$ et celui où le risque de première espèce est $\alpha = 0{,}05$?

2.4.5. Dans son texte, Gavarret dit qu'à la suite du tirage de 1000 boules et après l'observation de 900 blanches "personne n'hésiterait à affirmer que nécessairement les blanches étaient **beaucoup** plus nombreuses dans l'urne que les noires" (c'est nous qui soulignons). Or les tests proposés précédemment concernaient la question de confirmer ou infirmer la proposition : "nécessairement les blanches étaient plus nombreuses dans l'urne que les noires".

En l'absence de précision dans le texte sur l'interprétation de "beaucoup", on peut, parmi une infinité de possibles, proposer par exemple trois "interprétations" pour le mot "beaucoup" :

- Première interprétation : "beaucoup" signifie "il y a au moins trois fois plus de blanches que de noires dans l'urne" ;
- Deuxième interprétation : "beaucoup" signifie "il y a au moins quatre fois plus de blanches que de noires dans l'urne" ;
- Troisième interprétation : "beaucoup" signifie "il y a au moins neuf fois plus de blanches que de noires dans l'urne".

Pour chacune de ces "interprétations", en prenant toujours le même risque de première espèce que Gavarret, c'est-à -dire $\alpha = 0{,}0047$, construire la règle de décision, énoncer la décision qui aurait été prise à la suite du tirage de 1000 boules et après l'observation de 900 blanches et évaluer le "risque de deuxième espèce" à la suite du résultat.

Activité 2.5

À la fin de l'alinéa 1, Gavarret cherche en fait, à la suite du tirage de 1000 boules et après l'observation de 900 blanches, la valeur de ϵ tel que $|p - \widehat{p}| < \epsilon$. $\widehat{p}$, estimation de p, est fournie par la phrase : « sur dix boules sorties, on a amené neuf blanches et une noire » donc $\widehat{p} = 0{,}90$. N'ayant pas précisé avec quelle probabilité il cherche cet encadrement, on peut supposer que cette probabilité est celle qu'il attribue à des événements "certains" c'est-à -dire qu'elle est égale à $1 - \alpha$ où $\alpha = 0{,}0047$.

2.5.1. Contourner la « difficulté bien faite pour arrêter l'esprit le plus juste et le plus clairvoyant », en calculant ϵ et en déduire un intervalle de confiance pour p de niveau $0{,}9953$.

2.5.2. Gavarret ajoute encore une interrogation :

> **Essayons, en effet, de pénétrer plus avant, et demandons aux spectateurs combien il leur faudrait de tirages pour se croire fondés à regarder le résultat statistique comme représentant, *à très peu près* et dans des *limites données d'erreur possible*, la véritable composition de l'urne.**

Pour un ϵ donné, trouver la formule qui donne approximativement la taille de l'échantillon qui permettra d'obtenir $|p - \widehat{p}| \leq \epsilon$ avec une probabilité supérieure à $1 - 0{,}0047 = 0{,}9953$.

Activité 2.6

Dans l'alinéa 2, il considère que la proportion observée reste la même mais ceci à la suite d'un tirage de 20 boules.

On examine quatre cas correspondant à des interprétations de " beaucoup plus de blanches que de noires" (dans l'urne) :

- 1er cas : tester $p \leq 0{,}50$ contre $p > 0{,}50$;
- 2ème cas : tester $p \leq 0{,}75$ contre $p > 0{,}75$;
- 3ème cas : tester $p \leq 0{,}80$ contre $p > 0{,}80$;
- 4ème cas : tester $p \leq 0{,}90$ contre $p > 0{,}90$.

2.6.1. Calculer les coefficients binomiaux :

$$\binom{20}{20}, \qquad \binom{20}{19}, \qquad \binom{20}{18}, \qquad \binom{20}{17}, \qquad \binom{20}{16}.$$

2.6.2. On note Y la variable aléatoire qui, à un tirage de 20 boules dans une urne où la proportion de blanches est égale à $p = \dfrac{K}{N}$, associe le nombre y de blanches.

Pour $p = 0{,}50; 0{,}75; 0{,}80; 0{,}90$, calculer $\mathbb{P}_p((Y \geq y)$ avec $y = 16, 17, 18, 19$ et 20. Ici, dans la mesure où la taille de l'échantillon est inférieure à 30, on n'utilisera pas l'approximation normale mais la formule de la loi binomiale : $\mathbb{P}_p(Y \geq y) = \displaystyle\sum_{i=y}^{20} \binom{20}{i} p^i (1-p)^{20-i}$.

2.6.3. On note A_p le plus petit entier $y \leq n$ tel que $\mathbb{P}_p(Y \geq y) \leq 0{,}0047$.

ATTENTION. Si $\mathbb{P}_p(Y = n) > \alpha$, pour tout $y \leq n$ on a $\mathbb{P}_p(Y \geq y) > \alpha$ et il ne sera pas possible de trouver A_p. La probabilité de prendre la décision "$K > H$" sera toujours supérieure au "risque de première espèce" accepté et il ne sera pas possible de construire un test ayant α pour "risque de première espèce".

a) Pour chaque cas, en prenant toujours le même "risque de première espèce" que Gavarret, c'est-à -dire $\alpha = 0{,}0047$, construire la règle de décision, énoncer la décision qui aurait été prise à la suite du tirage de 20 boules et après l'observation de 18 blanches et quand il est possible de construire un test ayant un "risque de première espèce" $\alpha = 0{,}0047$, évaluer le "risque de deuxième espèce" à la suite du résultat.

b) Comparer alors l'évaluation de ce risque avec celui obtenu en tirant 1000 boules. Peut-on conjecturer que le "risque de deuxième espèce" diminue quand la taille de l'échantillon s'accroît ?

c) Prouver alors cette conjecture. Gavarret a-t-il raison en ne s'estimant pas suffisamment « éclairé par un si petit nombre d'épreuves » ?

2.6.4. Aujourd'hui, il est d'usage d'utiliser le "risque de première espèce" $\alpha = 0{,}05$ alors que Gavarret utilisait implicitement le risque $\alpha = 0{,}0047$.

1er problème : On souhaite toujours tester $H_0 : p \leq 0{,}50$ contre $H_1 : p > 0{,}50$. Supposons qu'à la suite du tirage de 1000 boules dans une urne, on ait tiré 532 blanches c'est-à -dire une proportion observée de $53{,}2\%$ de blanches.

Qu'en conclurait-on aujourd'hui ?

Qu'en aurait-on conclu avec le risque pris par Gavarret ?

2ème problème : On souhaite encore tester $H_0 : p \leq 0{,}50$ contre $H_1 : p > 0{,}50$. Supposons qu'à la suite du tirage de 2000 boules dans une urne, on ait tiré 1064 blanches c'est-à -dire encore une proportion observée de $53{,}2\%$ de blanches.

Qu'en conclurait-on aujourd'hui ?

Qu'en aurait-on conclu avec le risque pris par Gavarret ?

2.6.5. En gardant le même nombre $n = 1000$ de tirage dans l'urne et le même risque $\alpha = 0{,}0047$, 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 $H_1 : K < H$.

Donner les décisions obtenues à la suite de chacun des deux tests lorsque le nombre de boules blanches tirées est respectivement égal à 550, 502, 480 et 420 et comparer ces décisions.

Activité 3.

Cette activité représente une tentative d'illustration de la notion de chance variable.

Une population est composée de N individus numérotés $1, 2, ..., \alpha, ..., N-1, N$ possédant la pathologie donnée. L'événement "mort à la suite d'une pathologie pour l'individu numéro α" est un événement aléatoire dont la réalisation est assimilée au tirage d'une boule blanche dans une urne contenant K_α boules blanches et H_α noires.

Calculer la probabilité de l'événement : "l'individu numéro α meurt de la pathologie".

Calculer la probabilité de l'événement : "un individu tiré au hasard dans la population meurt de la pathologie".

Activité 4.

Compléter le tableau en ajoutant deux lignes :

- 6ème ligne avec la série initiale 600 morts sur 1500 malades.
- 7ème ligne avec la série initiale 840 morts sur 2100 malades.

Les mortalités moyennes en ajoutant la série 1 et en ajoutant la série 2 se rapprochent. Pouvait-on prévoir le résultat ?

Activité 5.

Dans les trois exemples précédents, comment se transformeraient les énoncés des résultats si, pour construire des intervalles de confiance, Gavarret avait utilisé ceux obtenus par les formules utilisées en classe de lycée et de B.T.S.

Quelle observation peut-on faire à partir de ces résultats concernant le rapport entre le degré de confiance à accorder à l'intervalle et la précision ?

Une application courante de la théorie des intervalles de confiance de niveau donné $1-\alpha$ consiste dans son utilisation pour tester l'hypothèse "lorsqu'un enfant naît, il y a autant de chance que ce soit un garçon qu'une fille" contre l'hypothèse "lorsqu'un enfant naît, les chances d'être un garçon ou une fille sont différentes" avec un "risque de première espèce" égal à α. On rejette l'hypothèse nulle si et seulement si $\frac{1}{2}$ n'appartient pas à l'intervalle construit.

A l'aide de cette règle, pour chacun des trois exemples, effectuer ce test, à la fois dans le cadre du "risque de premier espèce" pris par Gavarret et de celui utilisé dans les classes de lycée et de B.T.S.

Activité 6.

Pour le 3ème exemple du 1er groupe d'exemples (paragraphe 4.3), calculer la différence observée d et la limite l en utilisant la formule de Poisson. Obtient-on la même conclusion que Gavarret ?

Activité 7.

On note p_x la probabilité qu' un malade atteint d'une pathologie donnée soit guéri avec la dose x d'un médicament. On cherche à modéliser la relation entre p_x et x par $p_x = f(x)$ où f est une fonction donnée.

7.1. Dans un premier temps, on essaie la fonction f définie ainsi :
$f(x) = a + bx$ où a et b sont des coefficients réels.
Si on cherche une modélisation telle que la probabilité de guérir augmente avec la dose, quelle est la condition sur le coefficient b ?
Donner plusieurs exemples de valeurs pour a, b pour lesquels cette modélisation n'est pas adaptée.

7.2. La plupart du temps en médecine, les x mesurent log_{10}(dose) donc x peut prendre toutes les valeurs réelles.

a) Dans ce cas, montrer que les deux fonctions suivantes sont possibles pour une modélisation :

- Soit f définie par : $f(x) = \dfrac{e^{a+bx}}{1+e^{a+bx}}$
- Soit f définie par : $f(x) = F_Z(a+bx)$ où F_Z est la fonction de répartition de la loi normale réduite centrée.

b) Soient x_1 et x_2, deux log-doses différentes de la "médication". Pour x_1, on observe une proportion $\widehat{p_{x_1}}$ de personnes guéries et pour x_2, on observe une proportion $\widehat{p_{x_2}}$ de personnes guéries. D'après la "méthode numérique", on peut penser que $\widehat{p_{x_1}}$ estime bien p_{x_1} et que $\widehat{p_{x_2}}$ estime bien p_{x_2}.

On décide d'utiliser la première modélisation :

$$f(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$

Donner une méthode pour alors estimer les nombres inconnus a et b.

c) En déduire une solution au problème suivant :
Une dose de $d_1 = 2mg$ d'un remède est administrée à 20 patients. 12 sont guéris.
Une dose de $d_2 = 3mg$ du même remède est administrée à 25 autres patients. 18 sont guéris.
On ne peut augmenter la dose sans risques si elle est trop élevée.
Trouver la dose à administrer pour avoir une probabilité de guérison de 80%.

d) Qu'aurait dit Gavarret du résultat ?

7.3. Que feriez-vous si vous connaissiez les résultats $\widehat{p_{x_1}}$, $\widehat{p_{x_2}}$,...,$\widehat{p_{x_k}}$ pour k log-doses différentes x_1, x_2,...,x_k ?

Activité 8.

Compléter le tableau suivant :

Mortalité moyenne fournie par la statistique	STATISTIQUES de 950 cas
	Répartition des malades Erreur possible
	323 morts guéris

Énoncer la conclusion comme l'aurait fait Gavarret.

Activité 9.

Appliquer la "nouvelle formule" $l = 2\sqrt{\left(\frac{1}{\mu} + \frac{1}{\mu'}\right)\left(\frac{2m^*n^*}{\mu^{*2}}\right)}$ aux données de la 1ère version et aux données de la 2ème version et énoncer la conclusion. Pourquoi obtient-on avec cette nouvelle formule la même limite l pour les deux versions ?

Activité 10.

En utilisant la technique (erronée) qui consiste à calculer séparément pour p_1 et pour p_2 des intervalles de confiance de niveau $1 - 0{,}0047 = 0{,}9953$ et à décider que $p_1 \neq p_2$ si ces intervalles ne se chevauchent pas, quelles décisions auraient été prises concernant les trois versions de la page 23 du texte ? Comparer ces décisions à celles prises par Gavarret.

Activité 11.

Soit une population constituée de N personnes numérotées $\alpha = 1, 2, ..., N$. On note p_α la probabilité de contracter la maladie pour la personne α. Une mesure de santé publique est appliquée à l'ensemble de la population. On note p^*_α la probabilité de contracter la maladie pour la personne α après l'application de la mesure.

On pose $e_\alpha = p_\alpha - p^*_\alpha$, le gain de probabilité pour l'individu α.

Soient les deux situations suivantes avec $N = 4$ (données fictives ! ! !).

1ère situation :

p_α	0,9	0,6	0,8	0,9
p^*_α	0,6	0,5	0,4	0,9

2ème situation :

p_α	0,9	0,6	0,8	0,9
p^*_α	0,7	0,4	0,6	0,7

Calculer le gain en terme de "chance moyenne", obtenu par l'application de la mesure et comparer les deux situations.

Trouver une formule générale pour calculer le gain dans le cas N quelconque.

Dans le cas où $e_\alpha = e$ pour tout α, où e est le gain de probabilité identique pour chaque individu, y a-t-il des exemples où ce modèle est mis en défaut ?

Activité 12.

Pour comparer la saison chaude et humide avec la saison chaude et sèche, Gavarret a pris comme "rapport des fréquences" $\frac{n_1}{n_1 + n_2}$ où n_1 est le nombre correspondant à la deuxième ligne des données c'est-à -dire la saison chaude et humide mais pour comparer la saison chaude et humide avec la saison froide n_1 est le nombre correspondant à la première ligne des données.

Montrer qu'il est indifférent de prendre comme "rapport des fréquences" $\frac{n_1}{n_1 + n_2}$ ou $\frac{n_2}{n_1 + n_2}$.

Activité 13.

Justifier le test $H_0 : p =$ 4/7 contre $H_1 : p \neq$ 4/7.

Est-il équivalent à celui utilisé par Gavarret consistant dans l'artifice de multiplier n_1 par 3/4 et de tester ensuite $H_0 : p = 0{,}50$ contre $H_1 : p \neq 0{,}50$? Justifier mathématiquement la réponse.

Activité 14.

Calculer la valeur de la statistique de test et l'intervalle en utilisant l'artifice de multiplier n_1 par 3/4. Utiliser la "bonne méthode" n'utilisant pas l'artifice et conclure.

Activité 15.

On utilise trois médicaments M_1, M_2 et M_3 pour guérir une maladie et chaque médicament est testé sur un échantillon de malades. On pose p_i la proportion inconnue de personnes guéries par le médicament M_i, pour $i = 1, 2, 3$. Les données sont fictives.

1er médicament : 350 sont guéris dans un échantillon de 1000 personnes.

2ème médicament : 45 sont guéris dans un échantillon de 100 personnes.

3ème médicament : 420 sont guéris dans un échantillon de 1000 personnes.

Tester en utilisant la formule de Poisson :

1er médicament : $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.

2ère médicament : $H_0 : p_2 = p_3$ contre $H_1 : p_2 \neq p_3$.

3ère médicament : $H_0 : p_1 = p_3$ contre $H_1 : p_1 \neq p_3$.

Le résultat est-il contradictoire ?

Activité 16.

16.1. Dans l'exemple des individus atteints de dysenterie au Bengale, lors de la comparaison de la saison chaude et humide avec la saison chaude et sèche, il n'est pas indiqué le nombre d'habitants concernés pour chacune des saisons, mais on suppose cependant sans doute que c'est le même. En réalité, Gavarret teste l'égalité ou non de deux probabilités conditionnelles. En effet, on définit les événements suivants :

- S_1 : "être au Bengale pendant la saison chaude et humide" ;
- S_2 : "être au Bengale pendant la saison chaude et sèche" ;
- D : "être atteint de dysenterie".

On note :

- $p_1 = \mathbb{P}_D(S_1)$ probabilité d'être au Bengale dans la saison chaude et humide sachant qu'on est atteint de la dysenterie ;
- $p_2 = \mathbb{P}_D(S_2)$ probabilité d'être au Bengale dans la saison chaude et sèche sachant qu'on est atteint de la dysenterie ;
- $p_1^* = \mathbb{P}_{S_1}(D)$ probabilité d'être atteint de dysenterie sachant que l'on est au Bengale dans la saison chaude et humide ;
- $p_2^* = \mathbb{P}_{S_2}(D)$ probabilité d'être atteint de dysenterie sachant que l'on est au Bengale dans la saison chaude et sèche ;
- $p = \mathbb{P}(D)$ probabilité d'être atteint de dysenterie (cette probabilité est appelé par les médecins "prévalence de la dysenterie").

Le test effectué par Gavarret est donc en réalité le suivant :

$H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.
alors que le test intéressant est bien sûr :

$H_0 : p_1^* = p_2^*$ contre $H_1 : p_1^* \neq p_2^*$.
En effet, c'est celui qui permet de tester l'influence de la saison sur l'apparition de la dysenterie.

Montrer que $p_1 = p_2$ implique $p_1^* = p_2^*$ si et seulement si $\mathbb{P}(S_1) = \mathbb{P}(S_2)$.

16.2. Adopter la même démarche pour traiter les données de l'érysipèle et donner la conclusion.