Le test en statistique

Objectif

Comprendre la notion de test statistique, aborder des subtilités liées au choix de l'hypothèse initiale, parler d'estimateur, construire une zone de rejet, lire une table de gaussienne, manipuler des probabilités, voir des exemples.

I Introduction

La science du vivant s'intéresse à des milieux extrêmement complexes et bien souvent chaotiques, au sens mathématique du terme. La modification d'une seule donnée peut entraîner de vastes modifications en chaîne du système considéré. Pour apporter une réponse à une étude d'un macro-système comme le corps humain, il faut tenir compte d'un nombre incalculable de facteurs agissant au niveau microscopique. En ce sens une approche déterministe et mécanique est bien souvent insuffisante en biologie. Le fonctionnement d'un organe par exemple n'est pas du même niveau de complexité que le fonctionnement d'une horloge. Pour faire face à cette difficulté, la biologie s'est massivement tournée vers la statistique qui est le domaine d'application des mathématiques qui permet de traiter avec une grande rigueur des systèmes très complexes dont on ignore les paramètres et le fonctionnement mais sur lesquels on a accès à de nombreuses données expérimentales. C'est là tout le sens de la biostatistique et la raison pour laquelle le contenu mathématique de la PACES est essentiellement orienté vers cette branche appliquée des mathématiques.

L'objectif de ce cours est d'introduire la notion de test statistique ainsi que quelques principes et connaissances qui y sont rattachés. Cette présentation s'articulera sur l'exemple suivant.

Exemple 1. On considère une variété de souris et l'on suppose que la proportion de ces souris à développer un cancer est de 20%. On traite 100 souris par ce médicament et l'on observe que 15 d'entre elles ont développé malgré tout un cancer. A l'aide de cette étude, un organisme de contrôle de santé indépendant souhaite trancher si le médicament est efficace ou non sur cette population de souris.

II L'hypothèse nulle

A travers l'exemple que l'on vient de présenter on s'intéresse à la question suivante

« Le médicament est-il efficace ? »

Cependant notre étude étant statistique, elle va modéliser le chaos du système observé (la santé des souris) par un phénomène aléatoire. En considérant cette approche comme étant la négation d'une modélisation déterministe, les évènements observés ne seront pas certains, au sens probabiliste : les probabilités manipulées ne seront pas égales à 1. En conséquence il est tout à fait naturel que la réponse finale, quelle qu'elle soit, ne soit pas absolue. Pourtant notre approche étant mathématique (et donc rigoureuse !) nous devrons être capable de quantifier les erreurs commises, ce qui reviendra à donner une réponse avec un degré de confiance.

Néanmoins, avant de rentrer dans le traitement dur de notre étude, il nous faut auparavant la **modéliser**. Cette étape est cruciale et peut constituer une source considérable de confusions. Le mathématicien que je suis se permet d'insister. Les mathématiques sont d'une précision absolue lorsqu'elles sont correctement formulées. La modélisation quant à elle est toujours entachée d'erreur et la statistique dans ses applications est particulièrement sujette aux pièges où l'intuition peut être mise en défaut.

Pour répondre à notre question, nous allons tester la validité d'une hypothèse, par exemple

$$\textit{Le médicament ne permet pas de réduire l'apparition du cancer.} \qquad (H_0)$$

On appelle l'hypothèse que l'on considère l'hypothèse nulle (H_0) et sa négation, l'hypothèse 1 :

$$\textit{Le médicament permet de réduire l'apparition du cancer.} \qquad (H_1)$$

Il nous faudra alors décider entre deux possibilités :

1. rejeter l'hypothèse nulle (H_0) et accepter l'hypothèse 1 (H_1), ce qui dans notre exemple est plutôt positif, nous concluons que le médicament est probablement efficace ;
2. ne pas rejeter l'hypothèse nulle (H_0), ce qui dans notre exemple est plutôt décevant, nous concluons que le médicament n'est probablement pas efficace.

A chaque décision est associée une erreur :

1. rejeter l'hypothèse nulle, à tort c'est-à-dire alors qu'elle est vraie,
2. ne pas rejeter l'hypothèse nulle, à tort c'est-à-dire alors qu'elle est fausse.

Le problème est que pour diminuer le risque de commettre l'une des erreurs, il nous faudra augmenter le risque de commettre l'autre erreur. Cette impossibilité d'optimiser simultanément les deux erreurs provoque une dissymétrie dans les hypothèses.

Dans un test on préfère toujours par convention ne pas rejeter l'hypothèse nulle alors qu'elle est fausse (appelé erreur de seconde espèce) plutôt que de rejeter l'hypothèse nulle alors qu'elle est vraie (erreur de première espèce). Nous chercherons donc à minimiser la probabilité de rejeter l'hypothèse nulle à tort sans (dans ce cours) nous occuper de l'erreur de seconde espèce. En d'autres termes, notre angle d'attaque va être de *tout faire* pour ne pas rejeter l'hypothèse nulle (H_0). Cette phrase peut paraître anodine et peu rigoureuse, mais nous verrons plus tard dans les développements mathématiques comment elle se traduit. L'important pour l'instant est de bien comprendre la subtilité qu'entraîne une telle approche. Si malgré tous nos efforts pour ne pas rejeter l'hypothèse nulle (H_0) nous y sommes contraint par les résultats, notre conclusion est beaucoup plus forte. Dans notre exemple nous allons tout faire pour ne pas rejeter le fait que le médicament ne soit pas efficace et si malgré tous nos efforts nous sommes obligés de rejeter une telle hypothèse, il nous faut nous rendre à l'évidence, le médicament est très probablement efficace. Au contraire ne pas rejeter l'hypothèse nulle est une réponse beaucoup plus faible. Le médicament ne semble pas efficace : ce n'est peut-être pas si étonnant puisque nous avons fait notre possible pour favoriser cette hypothèse. Cette approche provoque un déséquilibre entre les deux hypothèses et voilà pourquoi l'hypothèse nulle est toujours celle que l'on ne souhaite pas démontrer.

Pour bien comprendre cette difficile notion, un bon exemple est celui de la justice française munie de la présomption d'innocence. Les deux erreurs sont les suivantes :

1. condamner un innocent
2. acquitter un coupable.

Dans ces deux erreurs, on préfère minimiser le risque de commettre la première quitte à augmenter celle de commettre la seconde. Ici l'hypothèse nulle est donc le fait que le suspect est innocent.

Si vraiment on est obligé de rejeter cette hypothèse alors on le considèrera coupable. Dans cette approche on condamne avec une plus de précision qu'on n'acquitte. Ainsi en statistique :

Définition II.1

- On appelle **hypothèse nulle** l'hypothèse sur laquelle on accepte plus facilement de se tromper. On note cette hypothèse ($\mathbf{H_0}$).
- L'hypothèse 1 notée ($\mathbf{H_1}$) est la négation de l'hypothèse nulle (H_0).

Proposition II.2

- Rejeter l'hypothèse nulle est un résultat plus fort que de ne pas rejeter l'hypothèse nulle.
- Afin de souligner la faiblesse du non-rejet, on ne dit pas que l'on *accepte* l'hypothèse nulle mais que l'on ne la rejette pas.

III Formulation mathématique

On relève l'état de santé de chaque souris traitée à la fin de l'expérience. On note 0 si la souris est saine et 1 si jamais elle a développé un cancer. On note n la taille de notre échantillon de souris, ici $n = 100$ et pour $i \in \{1, \dots n\}$, on suppose que l'état de la souris numéro i est donné par une variable aléatoire X_i. Naturellement chacune de ces variables aléatoires suit une loi de Bernoulli. On suppose que toutes les souris (traitées) ont la même probabilité de développer un cancer. On dit que les variables aléatoires sont identiquement distribuées. On note m cette probabilité :

$$\mathbb{P}(X_i = 1) = m.$$

Ce réel $m \in [0; 1]$ constitue la moyenne de la loi de Bernoulli :

$$\mathbb{E}(X_i) = m.$$

On suppose également que la probabilité pour une souris de développer un cancer est indépendant du fait que les autres souris développent ou non un cancer. En d'autres termes, on suppose que les variables aléatoires $X_1, \dots, X_n$ sont indépendantes. Cette hypothèse assez naturelle est très importante pour les considérations probabilistes qui suivent.

Hypothèses

On suppose que les variables aléatoires $(X_i)_{i\in\{1,\dots,n\}}$ sont

- indépendantes,
- identiquement distribués
- selon une loi de Bernoulli.

Remarque : dans la plupart des tests, l'hypothèse d'indépendance et l'hypothèse de même loi sont souvent vérifiées, cependant il n'est pas nécessaire que la loi soit celle de Bernoulli.

On sait que normalement les souris ont un probabilité $m_0 = 0,2$ de développer un cancer. Savoir si le traitement possède un effet bénéfique revient donc à savoir si la probabilité inconnue m des souris traitées de développer un cancer est plus élevée ou plus faible que cette proportion $m_0 = 0,2$. Plus précisément, les hypothèses (H_0) et (H_1) peuvent être reformulées de la façon suivante :

$$(m \geqslant m_0) \qquad (H_0)$$

et

$$(m < m_0) \qquad (H_1)$$

IV Les estimateurs

Un estimateur est une variable aléatoire qui « approche » un paramètre d'intérêt. Cette notion est volontairement floue et peut s'appliquer un peu près à n'importe quoi, on ne demande pas de qualité particulière dans la définition. Cependant, en contexte et en pratique, cette notion est naturelle.

Pour savoir dans notre cas si l'on rejette ou non l'hypothèse nulle (H_0), il nous faut mesurer ou plus exactement **approcher par l'expérience** notre moyenne m. Or un résultat phare des probabilités affirme que sous de bonnes hypothèses sur $(X_i)_{i\geqslant 1}$ (comme les nôtres), on a

$$\frac{X_1 + \cdots + X_n}{n} \underset{n\to+\infty}{\longrightarrow} m,$$

avec un sens pour la limite que je ne précise pas ici. Lorsque l'on fait une réalisation, une mesure concrète sur les souris, $x_1, x_2, \ldots, x_n$ des variables aléatoires $X_1, X_2, \ldots, X_n$, cela signifie que la moyenne *mesurée* $\frac{x_1+\cdots+x_n}{n}$ converge vers la moyenne *théorique* m. En ce sens, $\frac{X_1+\cdots+X_n}{n}$, noté $\bar{X_n}$ constitue un bon estimateur de notre paramètre m.

Proposition IV.1

Soit $(X_i)_{i\geqslant 1}$ une suite de variables aléatoires indépendante, identiquement distribuée et de moyenne $\mathbb{E}(X_i) = m$. Alors

$$\bar{X_n} = \frac{X_1 + \cdots + X_n}{n}$$

est un « bon » estimateur de m.

Donnons également un bon estimateur de la variance :

Proposition IV.2

Soit $(X_i)_{i\geqslant 1}$ une suite de variables aléatoires indépendante, identiquement distribuée de moyenne $\mathbb{E}(X_i) = m$ et de variance $\mathbb{E}\left(X_i^2\right) - \mathbb{E}\left(X_i\right)^2 = \sigma^2$. Alors

$$\hat{\sigma_n}^2 = \frac{\left(X_1 - \bar{X_n}\right)^2 + \cdots + \left(X_1 - \bar{X_n}\right)^2}{n}$$

est un « bon » estimateur de σ^2

V La zone de rejet

Puisque $\bar{X_n} = \frac{X_1+\cdots+X_n}{n}$ est un « bon » estimateur de m, on va construire un évènement (au sens probabiliste) $\mathscr{R}$, à partir de cette variable $\bar{X_n}$, sur lequel on décidera de rejeter l'hypothèse nulle.

Définition V.1

Une **zone de rejet** est un évènement probabiliste sur lequel nous déciderons de rejeter l'hypothèse nulle.

Dans notre exemple la zone de rejet devra être du type $\left\{\bar{X}_n \leqslant m_0\right\}$. Cependant puisque la taille de l'échantillon est finie, la variable aléatoire $\bar{X}_n$ approche m mais n'est pas rigoureusement égale à m. Plus précisément $\bar{X}_n$ est une variable aléatoire de moyenne m et plus n est grand plus $\bar{X}_n$ sera proche de m mais peut toujours fluctuer autour de sa moyenne. En conséquence, il est possible que $\bar{X}_n$ soit plus petit que m_0 alors que m lui soit plus grand que m_0. Dans ce cas nous serons dans la zone de rejet $\left\{\bar{X}_n \leqslant m_0\right\}$ à tort, nous rejetterons (H_0) alors qu'elle est vraie. On ne pourra jamais complètement éviter qu'une telle erreur se produise, mais on va la minimiser en diminuant notre zone de rejet :

$$\mathscr{R} = \left\{\bar{X}_n \leqslant m_0 - x_\alpha\right\},$$

où x_α est un réel que nous fixerons ultérieurement. En rétrécissant cette zone de rejet, nous diminuons la possibilité de rejeter à tort (H_0) mais nous augmentons la possibilité de ne pas rejeter à tort (H_0). Ce que l'on gagne d'un côté, on le perd de l'autre. Une dissymétrie dans les hypothèses est nécessaire pour contrôler l'erreur commise et par convention nous favorisons toujours la possibilité de ne pas rejeter à tort (H_0) contre la possibilité de rejeter à tort (H_0) (voir Proposition II.2).

Notre objectif est donc de minimiser la probabilité de l'évènement $\mathscr{R}$ sous l'hypothèse nulle. Cette probabilité est maximale lorsque l'on suppose que $m = m_0$. En effet, sous l'hypothèse nulle $m \geqslant m_0$. Or plus m est petit, plus $\bar{X}_n$ aura tendance à être petit et donc plus $\mathscr{R}$ aura une probabilité forte. On se place dans ce pire des cas en supposant $m = m_0$ et nous allons tenter de faire en sorte que cette probabilité $\mathbb{P}(\mathscr{R})$ soit plus petite qu'un seuil α. Par exemple $\alpha = 5\% = 0,05$.

Objectif

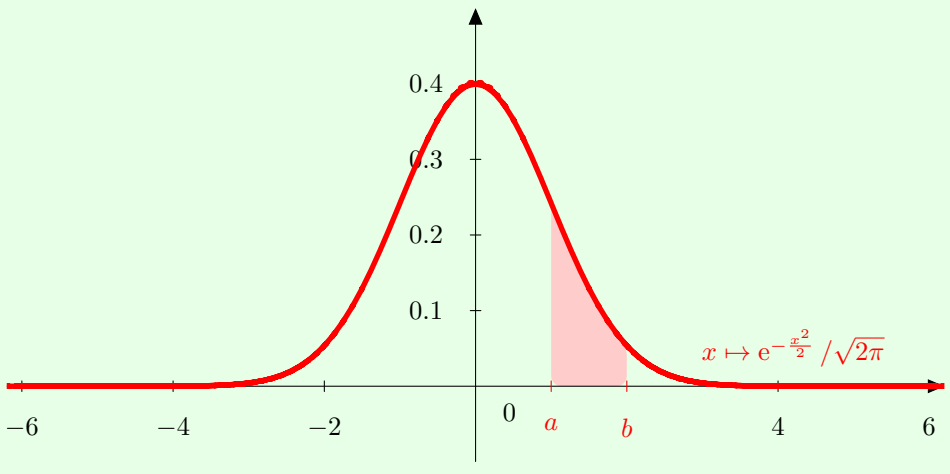
Trouver x_α pour que

$$\mathbb{P}(\mathscr{R}) = \mathbb{P}\left(\bar{X}_n > m_0 + x_\alpha\right) \leqslant \alpha.$$

VI La gaussienne et le Théorème Central Limite

Définition VI.1

Une **gaussienne** ou une **loi normale** classique est une variable aléatoire N qui peut prendre toutes les valeurs de $\mathbb{R}$. La probabilité d'avoir N entre l'intervalle $[a;b]$, avec a et b deux réels, est donnée par l'aire sous la courbe suivante entre a et b :



La loi normale est une loi centrale en théorie des probabilités du fait du résultat suivant.

Théorème VI.2 (Théorème Central Limite)

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendante et identiquement distribuée de moyenne $\mathbb{E}(X_1) = m_0$ et de variance $\mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \sigma^2$. Alors, la suite de variables aléatoires $(Z_n)_{n\geqslant 1}$ définie pour tout $n \geqslant 1$ par

$$Z_n = \frac{\frac{X_1+\cdots+X_n}{n} - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X_n} - m_0}{\frac{\sigma}{\sqrt{n}}},$$

converge quand n tend vers l'infini vers une loi normale N au sens suivant : pour tout $a \in \mathbb{R}$,

$$\lim_{n\to+\infty} \mathbb{P}(Z_n \leqslant a) = \mathbb{P}(N \leqslant a).$$

Ce résultat est fondamentale en probabilité et complète la Proposition IV.1. Non seulement $\bar{X_n}$ s'approche de m_0 mais ce théorème affirme que la vitesse d'approche est de l'ordre de $\sqrt{n}$:

$$\bar{X_n} - m_0 \approx \frac{\sigma}{\sqrt{n}} N.$$

Ainsi nous allons avoir accès à l'ordre de grandeur de l'erreur commise en approchant Z_n dans notre test par N.

Durant l'année de terminale, il est évoqué le théorème de Moivre-Laplace qui est un cas particulier de ce théorème dans le cas où la loi commune des X_n est une loi de Bernoulli. Le Théorème Central Limite énoncé (et admis !) ici est beaucoup plus puissant car n'impose aucune loi aux variables X_n. De cette façons la moyenne d'un grand nombre de phénomènes **inconnus** peut conduire (sous les hypothèses d'indépendance et de même distribution) à une variable aléatoire **connue** !

VII Ajustement de la zone de rejet

Lorsque n est petit et sous l'hypothèse d'avoir X_1 qui suit une loi de Bernoulli de paramètre m_0, on sait que $X_1 + \cdots + X_n$ suit une loi binomiale de paramètres n et m_0. Dans ce cas le calcul de x_α est immédiat à partir des tables bien connues des lois binomiales. Cependant lorsque n est grand il est préférable de considérer les tables de la gaussienne, approche qui est de toute façon indispensable lorsque la loi commune des X_i est inconnue. Approchons donc Z_n par N. Pour procéder à un tel amalgame peu rigoureux, il faut néanmoins justifier des **conditions de validité** suivantes : il faut que $n \geqslant 30$ que $np \geqslant 5$ et que $n(1-p) \geqslant 5$. Alors :

$$\begin{aligned}
\mathbb{P}(\mathscr{R}) &= \mathbb{P}\left(\bar{X_n} \leqslant m_0 - x_\alpha\right) \\
&= \mathbb{P}\left(\bar{X_n} - m_0 \leqslant -x_\alpha\right) \\
&= \mathbb{P}\left(\frac{\bar{X_n} - m_0}{\frac{\sigma}{\sqrt{n}}} \leqslant -\frac{x_\alpha}{\frac{\sigma}{\sqrt{n}}}\right) \\
&\simeq \mathbb{P}\left(N \leqslant -\frac{x_\alpha}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= \mathbb{P}\left(N \leqslant -\frac{\sqrt{n}x_\alpha}{\sigma}\right).
\end{aligned}$$

Notons que dans notre cas, σ^2, la variance de X_1, est bien connue : $\sigma^2 = m_0(1-m_0)$. Dans le cas d'une loi inconnue, il faut remplacer σ par son estimateur naturel $\hat{\sigma_n}$ construit dans la Proposition

IV.2. On obtient donc

$$\mathbb{P}(\mathscr{R}) \simeq \mathbb{P}\left(N \leqslant -\frac{\sqrt{n}x_\alpha}{\sqrt{m_0(1-m_0}}\right).$$

En remplaçant par les valeurs du problème :

$$\mathbb{P}(\mathscr{R}) \simeq \mathbb{P}\left(N \leqslant -\frac{10x_\alpha}{\sqrt{0,2\times 0,8}}\right).$$

Puis, **de tête**, $\sqrt{0,2\times 0,8} = \sqrt{0,16} = 0,4$ donc $\frac{10}{\sqrt{0,2\times 0,8}} = \frac{10}{0,4} = \frac{100}{4} = 25$. D'où,

$$\mathbb{P}(\mathscr{R}) \simeq \mathbb{P}(N \leqslant -25x_\alpha).$$

Supposons que l'on souhaite une précisions de 95%, cela revient à accepter une erreur de première espèce à un niveau de 5%. On veut donc $\mathbb{P}(\mathscr{R}) \leqslant 0,05$ c'est-à-dire

$$\mathbb{P}(N \leqslant -25x_\alpha) \leqslant 0,05.$$

Voici la table de la gaussienne :

	0	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7793	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8906	0,8925	0,8943	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Notez que l'on ne peut y lire que des valeurs positives. Cela provient du fait suivant. La loi normale est symétrique (voir le graphique de la définition VI.1) c'est-à-dire que pour tout réel $a \in \mathbb{R}$,

$$\mathbb{P}(N \leqslant -a) = \mathbb{P}(N \geqslant a)$$

et donc

$$\mathbb{P}(N \leqslant -a) = \mathbb{P}(N \geqslant a) = 1 - \mathbb{P}(N < a).$$

Ainsi,

$$\begin{aligned} \mathbb{P}(N \leqslant -25x_\alpha) \leqslant 0,05 \quad &\Leftrightarrow \quad 1 - \mathbb{P}(N < 25x_\alpha) \leqslant 0,05 \\ &\Leftrightarrow \quad 0,95 \leqslant \mathbb{P}(N < 25x_\alpha) \end{aligned}$$

Maintenant, à l'aide de la table, on en déduit que $25x_\alpha = 1,65$. Ainsi

$$x_\alpha = \frac{1,65}{25} = \frac{1,65}{5\times 5} = \frac{0,33}{5} = 0,066.$$

Finalement après tous ces développements, nous obtenons la version finale de notre zone de rejet :

$$\mathscr{R} = \left\{\bar{X}_n \leqslant m_0 - x_\alpha\right\} = \left\{\bar{X}_n \leqslant 0,134\right\}.$$

VIII Conclusion

Dans notre exemple, on nous affirme que 15 souris sur les 100 traitées ont été atteintes par un cancer. La proportion mesurée $\bar{x_n}$ est donc de $0,15$. Or

$$0,15 = \bar{x_n} \notin \mathscr{R}.$$

La proportion mesurée n'est pas dans la zone de rejet, nous ne rejetons donc pas (H_0). Malgré le fait que $0,15 \leqslant 0,2$ et que nous avions *l'impression* que le médicament était efficace, l'organisme de santé portera la conclusion :

« Avec une confiance de 95%*, nous ne rejetons pas le fait que le médicament soit inefficace. »*

Notez bien les subtilités. D'une part le résultat n'est valide qu'avec un degré de confiance. Il serait peut être différent avec un niveau de confiance plus faible. D'autre part l'organisme ne rejette pas l'hypothèse que le médicament soit inefficace, ce qui du fait de la dissymétrie des hypothèses, n'est pas la même chose qu'accepter que le médicament soit inefficace ou encore que de rejeter que le médicament soit efficace... Subtilité peu intuitive mais qui est une conséquence de notre approche.

Exemple 2. Pour se convaincre, appliquez à nouveau la méthode avec les mêmes valeurs mais du point de vue d'une entreprise pharmaceutique peu scrupuleuse qui préférerait se tromper en affirmant son médicament efficace que de se tromper en affirmant son médicament inefficace.

Il faut savoir qu'il existe toute une grande famille de tests utilisant différents résultats des probabilités et d'autres lois que la loi normale. Cependant l'approche présentée ici définit un schéma global que l'on retrouve dans de nombreux autres tests.

IX Exercices

Exercice 1. On souhaite savoir si le tabac a un effet sur le long terme sur le taux de triglycérides. Sur 120 patients ayant fumé pendant plus de dix ans, on note un taux moyen de 210 mg/dL avec un écart-type empirique de 50 mg/dL. Le taux moyen de triglycérides est usuellement de $1,3$ $g.L^{-1}$. Peut-on affirmer à 99% que le taux de triglycérides est anormalement élevé chez les fumeurs ?

Exercice 2. En France le pourcentage de personnes séropositives pour le VIH est de $0,23\%$. Une étude s'intéresse à un échantillon de 5000 personnes et note que 8 personnes dans cette étude ont été diagnostiquées séropositives. On souhaiterait savoir si cet échantillon confirme la valeur connue de $0,23\%$ ou non.

Exercice 3. On considère deux populations d'escargots de taille $n = 100$ chacune. On inocule un médicament à la population A et l'on observe que la taille moyenne des coquilles est de $21,2mm$ avec un écart type de $0,4$. On inocule le même médicament, mais sans son principe actif, à la population B. Cette fois la taille moyenne des coquilles est de $18,7$ avec un écart-type de $0,3$. Le médicament a-t-il une action sur la taille des coquilles de nos escargots ?