Annexe I

Théorie des tests :
Rappel très simplifié sur un exemple .

Test de l'efficacité d'un remède sur des malades atteint d'un rhume.

p_0 : probabilité de guérir dans les huit jours avec un placebo
p_1 : probabilité de guérir dans les huit jours avec le remède.

Deux hypothèses possibles :
- H_0 hypothèse nulle : $p_0 = p_1$ (pas d'effet du remède).
- H_1 hypothèse alternative : $p_0 \neq p_1$ (effet du remède)

Suite à une expérimentation où n_0 seront traités avec le placebo et n_1 avec le remède, observer le nombre de malades guéris : k_0 avec le placebo et k_1 avec le remède. A partir de cette observation, deux décisions seront possibles :
- D_0 : "H_0 est vraie" ou encore "pas d'effet du remède",
- D_1 : "H_1 est vraie" ou encore "effet du remède".

Démarche à utiliser :
- Construire une fonction qui dépend des observations et pour laquelle, quand l'hypothèse nulle est vraie, il est possible de calculer la loi de probabilité et ceci sans connaître les paramètres inconnus. Cette fonction appelée "statistique de test" sera notée φ ;
- Fixer α = risque de se tromper en décidant D_1 (c.a.d. "alors que H_0 est vraie"). C'est le risque dit "de première espèce" ;
- Déterminer une "zone critique" notée C_α telle que, lorsque l'hypothèse nulle est vraie, alors : $\mathbb{P}(\varphi \in C_\alpha) \leq \alpha$;
- Etablir la règle de décision suivante : rejet de H_0 (décision D_1) si et seulement si $\varphi \in C_\alpha$;
- Effectuer l'expérience et, suivant la valeur prise par φ, énoncer la décision. La construction du test garantit que la probabilité de se tromper en rejetant l'hypothèse nulle est inférieure à α.
- Evaluer le risque dit "de deuxième espèce". C'est le risque de se tromper en ne rejetant pas l'hypothèse nulle (décision D_0), c'est-à-dire $\mathbb{P}(\varphi \notin C_\alpha)$. Il est noté β.

Avant d'utiliser l'exemple, rappelons les notations suivantes :

Z désigne une variable aléatoire suivant une loi normale centrée réduite.
On note $Z \sim \mathcal{N}(0, 1)$.

F_Z désigne la fonction de répartition de la loi normale centrée réduite (souvent notée Π ou Φ), c'est-à-dire : $F_Z(z) = \mathbb{P}(Z \leq z)$ avec $Z \sim \mathcal{N}(0, 1)$.

z_α est le nombre réel tel que : $\mathbb{P}(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$ ou encore $\mathbb{P}(|Z| \geq z_\alpha) = \alpha$.
Remarque : $F_Z(z_{2\alpha}) = 1 - \alpha$.

Utilisation de la démarche :
- On pose :
  $f_0 = \frac{k_0}{n_0}$, $f_1 = \frac{k_1}{n_1}$ et $d = f_0 - f_1$;
  $f^* = \frac{k_0+k_1}{n_0+n_1}$, $s_d = \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}\sqrt{f^*(1 - f^*)}$ et $\varphi = \dfrac{d}{s_d}$
  On peut montrer que si $p_0 = p_1$, alors φ suit approximativement une loi normale centrée réduite $Z \sim \mathcal{N}(0, 1)$. Les conditions $n_0 \geq 30$, $n_1 \geq 30$, $n_0 p_0 \geq 5$, $n_0(1-p_0) \geq 5$, $n_1 p_1 \geq 5$ et $n_1(1-p_1) \geq 5$ permettent une bonne approximation.

- On prend $\alpha = 0{,}05$ c'est-à-dire 5% de chance de se tromper en disant que le remède fait de l'effet.
- On sait que $\mathbb{P}(|Z| > z_\alpha) = \alpha$ donc en prenant comme "zone critique", $C_\alpha =]-\infty, -z_\alpha[\cup]z_\alpha, +\infty[$, on a bien $\mathbb{P}(\varphi \in C_\alpha) \leq \alpha$. D'après les tables, $z_{0,05} = 1{,}96$, donc dire que la statistique appartient à la "zone critique" signifie que $\dfrac{|d|}{s_d} > 1{,}96$ ou encore que $|d| > l$ avec $l = 1{,}96 s_d$. On rejette l'hypothèse nulle si et seulement si $|d| > l$.
- Exécution du test.
  $n_0 = 200,\ \ k_0 = 140,\ \ f_0 = 0{,}70\ ,\ n_1 = 210,\ \ k_1 = 164,\ \ f_1 = 0{,}780\,952.$
  $d = -0{,}08095,\ \ s_d = 0{,}04325869,\ \ l = 1{,}96 \times 0{,}04325869 = 0{,}08478$
  $|d| = 0{,}08095 < 0{,}08478 = l$

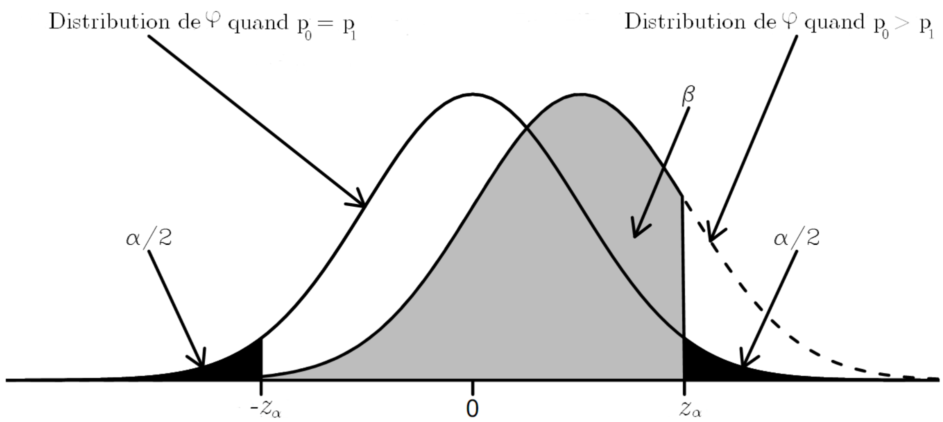
  Décision : D_0 c'est-à-dire non-rejet de H_0.
  On ne conclut pas à un effet du remède.
- Le risque de se tromper vaut $\beta = \mathbb{P}(\varphi \notin C_\alpha) = \mathbb{P}(-z_\alpha \leq \varphi \leq z_\alpha)$.
  Ici $\beta = \mathbb{P}(-1{,}96 \leq \varphi \leq 1{,}96)$.

  Dans le cas où on supposait $p_0 = p_1$ (hypothèse H_0), φ suivait loi normale centrée réduite, on avait alors $\mathbb{P}(-1{,}96 \leq \varphi \leq 1{,}96) = 1 - 0{,}05 = 0{,}95$.
  Pour calculer le "risque de deuxième espèce", il faut se placer dans le cas où on se trompe en décidant $p_0 = p_1$, donc (puisqu'on se trompe) dans le cas où $p_0 \neq p_1$ (hypothèse H_1). Mais alors dans ce cas, φ suit une loi de probabilité qui **dépend de** n_0, n_1, p_0, p_1 et cette loi **n'est pas la loi normale centrée réduite**.

  Pour un risque de première espèce donné α, le "risque de deuxième espèce" β dépend de n_0, n_1, p_0, p_1 et de α. Il n'est pas égal à α. Il peut être grand et même proche de $1 - \alpha$ (et ici proche de $0{,}95$!).

  Pour cette raison, lorsque la décision D_0 est prise, il faut, quand c'est possible, assortir cette décision de l'évaluation du "risque de deuxième espèce", et sinon, il est alors préférable de dire qu'on ne rejette pas l'hypothèse nulle. Dans le cas de ce test de comparaison entre deux probabilités, l'expression exacte de ce "risque de deuxième espèce" n'est pas donnée dans cette annexe car trop complexe mais il existe des tests (voir Activités 2) où elle peut être exprimée aisément.

Les deux risques



Annexe II

Intervalle de confiance.

Rappel très simplifié sur un exemple.

p proportion inconnue de patients ayant eu un effet secondaire suite à l'absorption d'un remède. Un échantillon de n patients sera observé. Sur ces n patients, k ont eu un effet secondaire.

Démarche à effectuer :

- Déterminer le niveau de confiance $1-\alpha$ avec $0<\alpha<1$.
- Trouver deux fonctions φ_1 et φ_2 de k (donc pour lesquelles le résultat sera aléatoire) telles que $\mathbb{P}(\varphi_1 \leq p \leq \varphi_2) \geq 1-\alpha$.
- A la suite de l'expérience, appliquer les deux fonctions à k. L'intervalle de confiance de niveau $1-\alpha$ pour p sera $[\varphi_1(k)\,,\,\varphi_2(k)]$.

On pose :

$$\varphi(k) = \frac{\dfrac{k}{n}-p}{\sqrt{\dfrac{\dfrac{k}{n}(1-\dfrac{k}{n})}{n}}},$$

On peut démontrer que, si les conditions suivantes sont réalisées : $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, alors φ suit approximativement une loi normale centrée réduite notée $\mathcal{N}(0,1)$.

Si Z suit une loi normale centrée réduite, les tables permettent de calculer z_α vérifiant :
$\mathbb{P}(-z_\alpha \leq Z \leq z_\alpha) = 1-\alpha$.

Donc $\mathbb{P}(-z_\alpha \leq \varphi \leq z_\alpha) = 1-\alpha$. En remplaçant φ par son expression, on trouve :

$\mathbb{P}(f-e \leq p \leq f+e) = 1-\alpha$ avec $f = \dfrac{k}{n}$ et $e = z_\alpha\sqrt{\dfrac{\dfrac{k}{n}(1-\dfrac{k}{n})}{n}} = z_\alpha\sqrt{\dfrac{f(1-f)}{n}}$.

On aura ici $\varphi_1(k) = f-e$ et $\varphi_2(k) = f+e$.

L'intervalle de confiance de niveau $1-\alpha$ pour p sera $[f-e\ ,\ f+e]$.

Dans un échantillon de $n = 1100$ patients ayant absorbé le remède, 583 ont eu un effet secondaire. On cherche un intervalle de confiance de niveau $0{,}95$ pour p.

Ici $z_{0{,}95} = 1{,}96$

$f = 0{,}53$ et $e = 1{,}96\sqrt{\frac{f(1-f)}{n}} = 0{,}0294$

Ici l'intervalle de confiance est : $[0{,}5006\,;0{,}5594]$

Programme B.T.S. : $[f-\frac{1}{\sqrt{n}}\ ,\ \ f+\frac{1}{\sqrt{n}}]$ intervalle de confiance de niveau $0{,}95$ pour p.

Ici $[0{,}49985\,;0{,}56015]$ intervalle de confiance de niveau $0{,}95$ pour p.

Annexe III

Une démonstration actuelle est la suivante :

Soit p la probabilité de réalisation de E et f_n la fréquence observée de l'événement E à la suite de n expériences indépendantes . On pose :

$Y_n = \dfrac{f_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$, $T_n = \sqrt{\dfrac{p(1-p)}{f_n(1-f_n)}}$, $U = 1$ variable aléatoire constante.

$Y_nT_n = \dfrac{f_n - p}{\sqrt{\dfrac{f_n(1-f_n)}{n}}}$

Z une variable aléatoire de loi normale $\mathcal{N}(0,1)$

D'après le théorème de Moivre-Laplace :

$Y_n \xrightarrow{\mathcal{L}} Z$ (convergence en loi vers la loi normale).

Mais :

$f_n \xrightarrow{p.s.} p$ (convergence presque sûre d'après la loi forte des grands nombres)

Donc $T_n \xrightarrow{p.s.} U$

Puisque U est une variable aléatoire constante :

$Y_nT_n \xrightarrow{\mathcal{L}} ZU$ donc $Z_n \xrightarrow{\mathcal{L}} Z$.

Ceci entraîne :

$$\mathbb{P}\left(-z_\alpha \leq \frac{f_n - p}{\sqrt{\dfrac{f_n(1-f_n)}{n}}} \leq z_\alpha\right) \longrightarrow \mathbb{P}(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$$

Ou encore :

$$\mathbb{P}\left(f_n - z_\alpha\sqrt{\frac{f_n(1-f_n)}{n}} \leq p \leq f_n + z_\alpha\sqrt{\frac{f_n(1-f_n)}{n}}\right) \simeq 1 - \alpha$$

Il est alors facile de voir que ce résultat est aussi celui de Gavarret.

Posons : $\alpha = \dfrac{2}{\sqrt{\pi}}\displaystyle\int_u^\infty e^{-t^2}dt$

Le changement de variable $t = \dfrac{x}{\sqrt{2}}$ donne :

$\dfrac{\alpha}{2} = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{u\sqrt{2}}^{+\infty} e^{-\frac{x^2}{2}}dx = 1 - F_Z(u\sqrt{2})$ où F_Z est la fonction de répartition de la loi normale centrée réduite.

Donc $u\sqrt{2} = F_Z^{-1}(1 - \dfrac{\alpha}{2}) = z_\alpha$

Gavarret pose : $s = u\sqrt{\frac{2.m.n}{\mu^3}}$, ce qui lui donne l'intervalle de confiance de niveau P pour B de la forme : $[\frac{m}{\mu} - s,\ \frac{m}{\mu} + s]$.

Puisque $\frac{m}{\mu} = f_n$, $\frac{n}{\mu} = 1 - f_n$ et on a bien $s = z_\alpha\sqrt{\dfrac{f_n(1-f_n)}{n}}$. En prenant $\alpha = 0{,}0047$ (donc $P = 0{,}9953$), $z_\alpha \simeq 2\sqrt{2}$, l'intervalle de confiance de niveau $1-\alpha$ pour p (ici égal à B) est bien le même que celui annoncé par Gavarret.

Annexe IV

Une démonstration actuelle possible utilise une généralisation du théorème central limite.

On pose $X_i^{(1)} \sim \mathcal{B}(p_1)$ pour $1 \leq i \leq n_1$ (resp. $X_k^{(2)} \sim \mathcal{B}(p_2)$ pour $1 \leq k \leq n_2$) la variable qui vaut 1 si E est arrivé dans la première série (resp. dans la deuxième série)

Soit $f_{n_1} = \dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} X_i^{(1)}$ (resp. $f_{n_2} = \dfrac{1}{n_2}\displaystyle\sum_{k=1}^{n_2} X_k^{(2)}$) la fréquence d'apparition de E dans la première série (resp. la deuxième série).

Soit $Y_j = \dfrac{1}{n_1} X_j^{(1)}$ si $1 \leq j \leq n_1$ et $Y_j = -\dfrac{1}{n_2} X_{j-n_1}^{(2)}$ si $n_1 + 1 \leq j \leq n_1 + n_2$.

On pose encore :

$$W_{(n_1,n_2)} = \frac{Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2} - \mathbb{E}(Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2})}{\sqrt{var(Y_1 + \cdots + Y_{n_1} + Y_{n_1+1} + \cdots + Y_{n_1+n_2})}}$$

Les variables Y_j n'ont pas toutes la même loi mais sous des conditions dites de Liapounov qui sont ici vérifiées, on a encore :

$$W_{(n_1,n_2)} \xrightarrow{\mathcal{L}} Z$$

où Z est une variable aléatoire de loi $\mathcal{N}(0,1)$.

Mais :

$$W_{(n_1,n_2)} = \frac{(f_{n_1} - f_{n_2}) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

Si on pose :

$$T_{(n_1,n_2)} = \frac{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}{\sqrt{\dfrac{f_{n_1}(1-f_{n_1})}{n_1} + \dfrac{f_{n_2}(1-f_{n_2})}{n_2}}}$$

D'après la loi forte des grands nombres :

$$T_{(n_1,n_2)} \xrightarrow{p.s.} U$$

où U est la variable aléatoire constante égale à 1.

Alors :

$$W_{(n_1,n_2)} T_{(n_1,n_2)} \xrightarrow{\mathcal{L}} Z.U = Z$$

et donc :

$$Z_{(n_1,n_2)} = \frac{(f_{n_1} - f_{n_2}) - (p_1 - p_2)}{\sqrt{\dfrac{f_{n_1}(1-f_{n_1})}{n_1} + \dfrac{f_{n_2}(1-f_{n_2})}{n_2}}} \xrightarrow{\mathcal{L}} Z$$

Sous l'hypothèse nulle, $p_1 = p_2$ et par conséquent :

$$\mathbb{P}\left(|f_{n_1} - f_{n_1}| \leq z_\alpha \sqrt{\frac{f_{n_1}(1-f_{n_1})}{n_1} + \frac{f_{n_2}(1-f_{n_2})}{n_2}}\right) \simeq 1 - \alpha$$

En reprenant $\alpha = 0{,}0047$, on retrouve la formule de Gavarret.