Université de Picardie Jules Verne **2017-2018**
UFR des Sciences

Master 1 2IBS - Semestre 1
Biostatistique

Echantillonnage, estimation, intervalle de confiance, test statistique
Cas d'une ou de deux proportions

1. Simulations

1.1. Loi de Bernoulli et simulation

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé.

Une variable aléatoire X suit **la loi de Bernoulli** de paramètre $p \in]0,1[$, que l'on note $\mathcal{B}(p)$, si et seulement si X est à valeurs dans $\{0;1\}$, et $P(X=1)=p$ et $P(X=0)=1-p$.

Une telle variable aléatoire permet d'indiquer si un événement A est réalisé ($X=1$) ou pas ($X=0$). Comme exemples d'application on peut citer :

- lancer d'une pièce menant à Pile ou Face, A = "obtenir Pile" ;
- tirer une boule dans une urne contenant des boules blanches et noires, A = "obtenir une blanche" ;
- choisir d'un individu dans la population, A = "l'individu est malade".

Ainsi , une telle variable **permet de représenter un caractère qualitatif à deux modalités**.

Simulation 1

p étant donné dans $]0,1[$, on considère une urne contenant une proportion p de boules blanches. Plus précisément, on considère l'entier N plus petit multiple de 10 tel que Np soit entier, et ainsi une urne contenant N boules, dont Np boules blanches et $N(1-p)$ boules noires. Par exemple, pour $p=0,42$, on a $N=100$, $Np=42$ et $N(1-p)=58$.

On suppose que les N boules sont numérotées de 1 à N, de 1 à Np pour les boules blanches, de $Np+1$ à n pour les noires.

A l'expérience aléatoire " tirer une boule au hasard dans l'urne ", on peut associer l'univers $\Omega=\{1,\ldots,N\}$ et le munir de l'équiprobabilité P.

Dans ce contexte, l'événement A "obtenir une boule blanche" est $A=\{1,\ldots,Np\}$, sa probabilité étant alors $P(A)=\frac{cardA}{card\Omega}=\frac{Np}{N}=p$.

Considérant la variable aléatoire X qui à chaque tirage d'une boule associe 1 si elle est blanche et 0 sinon, on a $(X=1)=A$ et $(X=0)=\overline{A}$, et donc $P(X=1)=P(A)=p$ et $P(X=0)=P(\overline{A})=1-P(A)=1-p$.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 1)

Le tirage d'une boule de l'urne est simulé par l'instruction =ALEA.ENTRE.BORNES(1;*N*) à entrer dans la cellule B8 (par exemple).

La valeur correspondante de X est alors obtenue par l'instruction =SI(B8<=*Np*;1;0).

Simulation 2

A l'expérience aléatoire "choisir un nombre au hasard dans l'intervalle $[0;1]$" on peut associer une variable aléatoire Y suit la loi Uniforme sur l'intervalle $[0;1]$ (loi à densité) ; Y indique le nombre obtenu. On sait que pour tout $y \in [0;1]$, $P(Y \leq y)=y$.

p étant donné dans $]0,1[$, on a alors $P(Y \leq p)=p$. Considérant la variable aléatoire X définie par $(X=1)=(Y \leq p)$ et $(X=0)=\overline{(Y \leq p)}=(Y>p)$, X suit la loi de Bernoulli $\mathcal{B}(p)$.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 2)

Une valeur de Y est simulée par l'instruction =ALEA() à entrer dans la cellule B7 (par exemple).

La valeur correspondante de X est alors obtenue par l'instruction =SI(B7<=*p*;1;0).

1.2. Loi binomiale et simulation

Reprenons l'exemple d'une urne contenant une proportion $p = 0,42$ de boules blanches.

On tire une boule au hasard dans l'urne : le nombre de "boule blanche" obtenu en un tirage est une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$: $P(X = 1) = p = 0.42$ et $P(X = 0) = 1 - p = 0,58$. On a $E(X) = p = 0,42$ et $Var(X) = p(1 - p) = 0,2436$.

Si on effectue $n = 50$ tirages avec remise d'une boule, on observe la réalisation de X_1 , X_2 , ... , X_{50} , variables aléatoires indépendantes de même loi que X. On dit que l'on a un échantillon aléatoire simple de taille $n = 50$ de loi de Bernoulli de paramètre $p = 0,42$.

La proportion de "boules blanches" obtenue est une variable aléatoire :

$$F_n = \frac{X_1 + X_2 + \cdots + X_{50}}{50} = \frac{\sum_{i=1}^{n} X_i}{n}$$

où $\sum_{i=1}^{n} X_i$ représente le nombre de "boules blanches" obtenues en $n = 50$ tirages.

Ayant procédé par répétitions d'expériences indépendantes, $nF_n = \sum_{i=1}^{n} X_i$ est une variable aléatoire de la loi Binomiale $\mathcal{B}(50; 0,42) = \mathcal{B}(n, p)$.

On a donc $nE(F_n) = E(nF_n) = np$ et $n^2 Var(F_n) = Var(nF_n) = np(1 - p)$, d'où $E(F_n) = p = 0,42$ et $Var(F_n) = \frac{p(1-p)}{n} = \frac{0,2436}{n}$.

On constate donc que lorsqu'on augmente la taille n de l'échantillon, l'espérance de F_n reste constante, égale à $0,42$, alors que la variance diminue.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 1 et 2)

On reprend les simulations 1 et 2 en répétant 50 les instructions précédentes sur 50 lignes. Il suffit ensuite de "sommer" les valeurs de X obtenues pour avoir le nombre de boules blanches obtenues, puis de diviser par 50 pour avoir la fréquence.

2. Echantillonnage : cas d'une proportion

2.0. Quel cadre mathématique ?

Statistique et probabilités :

Description des observations et modèle théorique.

La Statistique consiste à étudier un ensemble d'objets (on parle de population, composée d'individus ou unités statistiques) sur lesquels on observe des caractéristiques, appelées variables statistiques.

Le calcul des Probabilités permet de proposer un modèle théorique d'une situation concrète afin de quantifier la fiabilité des affirmations.

Population et échantillon :

Dans certains cas on peut obtenir les valeurs de ces variables sur l'ensemble de la population ; en appliquant les méthodes de la statistique descriptive il est possible, au moyen de tableaux, graphiques, paramètres, d'analyser ces résultats. Exemples : Recensement de la population française, notes obtenues par tous les candidats à un examen, salaires de tous les employés d'une entreprise, etc...

Mais la population peut être trop vaste pour être étudiée dans sa totalité, par manque de moyens, ou de temps. (C'est le cas si on s'intéresse aux intentions de vote des Français pour une élection). Elle peut même être considérée comme infinie. C'est le cas si l'on note la qualité (défectueuse ou non) des pièces produites par un certain procédé : le nombre de ces pièces est a priori illimité, et on ne peut toutes les tester.

De même, si l'on s'intéresse aux fréquences d'obtentions de "pile" et "face" avec une pièce de monnaie, le nombre de lancers de pièce à étudier est a priori infini : on a ici une population latente infinie.

Il arrive aussi que la mesure d'une variable soit destructrice pour l'individu : si on étudie la durée de vie de certains appareils, il serait absurde de les faire tous fonctionner jusqu'à la panne, les rendant inutilisables.

Dans tous ces cas, on est amené à n'étudier qu'une partie de la population, un échantillon, obtenu par sondage, dans le but d'extrapoler à la population entière des observations faites sur l'échantillon.

Fluctuation d'échantillonnage

Lorsqu'on étudie un caractère sur plusieurs échantillons d'une même population, on peut observer que les résultats ne sont pas identiques selon les échantillons. Plus la taille de l'échantillon étudié est grande, plus les résultats obtenus seront fiables. Cela s'explique par la diminution de la variance, et aussi par la loi des grands nombres.

La fluctuation d'échantillonnage représente la fluctuation entre les différents résultats obtenus d'une même enquête sur différents échantillons d'une même population.

Ces différents résultats présentent une certaine régularité, ce qui se traduit par la notion d'intervalle de confiance.

2.1. Caractère statistique et variable aléatoire

Considérons une population Ω sur laquelle on définit un caractère qualitatif à deux modalités A et B. On convient de représenter la modalité A par 1 et la modalité B par 0.

Le caractère est ainsi représenté par une application X de Ω dans $\mathbb{R}$ qui, à tout individu ω, associe un réel $x = X(\omega) \in X(\Omega) = \Omega_X = \{0, 1\}$ ensemble des "valeurs" du caractère.

Cette application modélise le caractère de façon déterministe : si on connaît l'individu ω, on connaît aussitôt la valeur x. Son étude relève de la statistique descriptive qui conduit, par exemple, au tableau des couples (x_i, f_i) où x_i est une valeur observée et f_i sa fréquence.

Supposons maintenant que l'on tire au hasard un individu ω dans cette population Ω pour consigner la valeur x du caractère. Ne pouvant pas prévoir quel individu précis sera tiré, on ne peut pas prévoir non plus la valeur précise de x qui sera consigner. On aimerait donc disposer d'un moyen d'attribuer une probabilité aux éléments de Ω_X.

Ici, X est une variable aléatoire de loi de Bernoulli $\mathcal{B}(p)$ où p est la proportion d'individus ayant la modalité A dans la population : $P(X = 1) = p$ et $P(X = 0) = 1 - p$.

2.2. Echantillonnage

Lorsqu'on n'a pas accès à l'ensemble de la population, la proportion p est inconnue. On procède à un **échantillonnage**, i.e. au choix de n individus dans la population, sur lesquels on observe la valeur x du caractère X. Lorsque les tirages ont lieu avec (respectivement sans) remise, l'échantillonnage est dit non-exhaustif (resp. exhaustif). Lorsque la taille n de l'échantillon est faible par rapport à celle N de la population ($N \geq 10n$), alors tout échantillonnage est assimilable au cas non-exhaustif.

Pour un premier échantillonnage, on observera des valeurs x_1, x_2, ..., x_n du caractère. Pour un deuxième échantillonnage de même taille, on observera des valeurs x'_1, x'_2, ..., x'_n du caractère. Et ainsi de suite. On peut alors considérer la suite x_1, x'_1, ... comme les valeurs observées d'une même variable aléatoire X_1, la suite x_2, x'_2, ... comme les valeurs observées d'une même variable aléatoire X_2, ... Ainsi, pour tout $i = 1, \ldots, n$, la variable aléatoire X_i correspond aux valeurs du caractère du i-ème individu obtenu par échantillonage, et aura donc la **même loi de probabilité que** X. De plus, l'échantillonnage étant non-exhaustif (tirages avec remise), les variables aléatoires X_i sont indépendantes.

Plus précisément, les variables aléatoires X_i sont des applications de Ω^n dans $\mathbb{R}$, qui à tout échantillonnage $(\omega_1, \omega_2, \ldots, \omega_n)$ associe $x_i = X_i(\omega_1, \omega_2, \ldots, \omega_n) = X(\omega_i)$

On dira que $(X_1, X_2, \ldots, X_n)$ est un **échantillon** (aléatoire simple) **de taille** n **de** X, et que $(x_1, x_2, \ldots, x_n)$ est une observation de l'échantillon.

Le terme d'échantillon désigne à la fois les n individus choisis et le n-uple de variables aléatoires $(X_1, X_2, \ldots, X_n)$.

2.3. Estimateur et estimation d'une proportion

Objectif : déterminer p à l'aide d'informations obtenues à partir d'un échantillonnage de taille n extrait de la population. Impossible tant que $n < N$, mais la théorie de l'échantillonnage conduit à des **estimations** $\widehat{p}$ de p, d'autant meilleures que n est grand.

Estimateur du paramètre p : **proportion** (ou fréquence) **d'échantillon** $F_n = \frac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre d'individus de l'échantillonnage ayant la modalité A.

Pour une observation $(x_1, x_2, \ldots, x_n)$ de l'échantillon (en pratique on observe souvent directement $\sum_{i=1}^{n} x_i$), une **estimation ponctuelle** de p est $f_n = \frac{\sum_{i=1}^{n} x_i}{n} = \hat{p}$.

2.4. Proportion d'échantillon

Un exemple sur la proportion

Un groupe de 4 enfants, Alexis, Benjamin, Cyril et David, d'âges respectifs 12, 13, 14 et 15 ans.

On choisit un enfant au hasard dans le groupe, on peut considérer :

- X, indicatrice du fait que l'enfant plus 14,5 ans,

  variable aléatoire de loi de Bernoulli $\mathcal{B}\left(\frac{1}{4}\right)$:

  $P(X = 1) = \frac{1}{4} = p$ et $P(X = 0) = \frac{3}{4} = 1 - p$.

Cherchons à retrouver ou à approcher ces résultats à partir d'échantillons non-exhaustifs (**avec remise**) de taille $n = 3$. Il y en a $4^3 = 64$, ils forment un univers Ω', ensemble des résultats possibles de l'expérience aléatoire "choisir un échantillon".

On peut munir Ω' de la tribu des événements $\mathcal{A}' = \mathcal{P}(\Omega')$ et de l'équiprobabilité P' sur $(\Omega', \mathcal{A}')$. A chacun des résultats (échantillons) ω, on peut associer la proportion $F_n(\omega) = f_n$ d'enfants ayant plus de 14,5 ans. On obtient les résultats suivants :

ω	f_n	ω	f_n	ω	f_n	ω	f_n
(A,A,A)	0	(B,A,A)	0	(C,A,A)	0	(D,A,A)	1/3
(A,A,B)	0	(B,A,B)	0	(C,A,B)	0	(D,A,B)	1/3
(A,A,C)	0	(B,A,C)	0	(C,A,C)	0	(D,A,C)	1/3
(A,A,D)	1/3	(B,A,D)	1/3	(C,A,D)	1/3	(D,A,D)	2/3
(A,B,A)	0	(B,B,A)	0	(C,B,A)	0	(D,B,A)	1/3
(A,B,B)	0	(B,B,B)	0	(C,B,B)	0	(D,B,B)	1/3
(A,B,C)	0	(B,B,C)	0	(C,B,C)	0	(D,B,C)	1/3
(A,B,D)	1/3	(B,B,D)	1/3	(C,B,D)	1/3	(D,B,D)	2/3
(A,C,A)	0	(B,C,A)	0	(C,C,A)	0	(D,C,A)	1/3
(A,C,B)	0	(B,C,B)	0	(C,C,B)	0	(D,C,B)	1/3
(A,C,C)	0	(B,C,C)	0	(C,C,C)	0	(D,C,C)	1/3
(A,C,D)	1/3	(B,C,D)	1/3	(C,C,D)	1/3	(D,C,D)	2/3
(A,D,A)	1/3	(B,D,A)	1/3	(C,D,A)	1/3	(D,D,A)	2/3
(A,D,B)	1/3	(B,D,B)	1/3	(C,D,B)	1/3	(D,D,B)	2/3
(A,D,C)	1/3	(B,D,C)	1/3	(C,D,C)	1/3	(D,D,C)	2/3
(A,D,D)	2/3	(B,D,D)	2/3	(C,D,D)	2/3	(D,D,D)	1

On définit ainsi une variable aléatoire F_n, dont on peut obtenir la loi de probabilité :

x_i	0	1/3	2/3	1
$P(F_n = x_i)$	27/64	27/64	9/64	1/64

On peut alors calculer :

- $E(F_n) = \frac{1}{4}$: on remarque que $E(F_n) = p = E(X)$.
- $Var(F_n) = \frac{1}{16}$: on remarque que $Var(F_n) = \frac{p(1-p)}{n} = \frac{Var(X)}{n}$.

Propriétés générales de $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$.

$nF_n = \sum_{i=1}^{n} X_i$ suit la loi Binomiale $\mathcal{B}(n,p)$. On a alors $nE(F_n) = E(nF_n) = np$ et $n^2 Var(F_n) = Var(nF_n) = np(1-p)$, d'où $E(F_n) = p$ et $Var(F_n) = \dfrac{p(1-p)}{n}$.

On a ainsi $E(F_n) = p$ et on dit que F_n est un **estimateur sans biais** de p.

On a de plus $\lim_{n\to+\infty} Var(F_n) = 0$ et on dit que F_n est un **estimateur convergent** de p.

Théorème. **Loi faible des grands nombres**

Si les X_i sont indépendantes et admettent la même espérance p et la même variance σ^2,
alors pour tout $\varepsilon > 0$, $\lim_{n\to+\infty} P(|F_n - p| > \varepsilon) = 0$; cette convergence étant uniforme en p.
Cela signifie que (F_n) converge en probabilité vers p.

Théorème central limite

Si les X_i sont indépendantes, de même espérance mathématique μ et de même écart-type σ,
et si $\overline{X_n} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, alors $Z_n = \dfrac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$;

autrement dit que $\overline{X_n}$ suit approximativement la loi normale $\mathcal{N}\left(\mu; \dfrac{\sigma}{\sqrt{n}}\right)$.

De plus, si $np \geq 10$ et $n(1-p) \geq 10$, on peut approcher la loi Binomiale $\mathcal{B}(n,p)$ par la loi normale $\mathcal{N}\left(np\ ; \sqrt{np(1-p)}\right)$. On en déduit que nF_n suit approximativement la loi normale $\mathcal{N}\left(np\ ; \sqrt{np(1-p)}\right)$, et donc F_n suit approximativement la loi normale $\mathcal{N}\left(p\ ; \sqrt{\dfrac{p(1-p)}{n}}\right)$. Ainsi, $U = \dfrac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$.

Commentaires de ces résultats

F_n a toujours pour espérance p : la proportion dans l'échantillon est, "en moyenne", celle de la population.

La variance de F_n est d'autant plus faible que n est grand : la proportion dans l'échantillon varie d'autant moins d'un échantillon à l'autre que la taille de cet échantillon est grande.

A la limite, si n tend vers l'infini, $Var(F_n)$ tend vers 0 et donc F_n tend vers la constante p.

Dans la pratique, l'approximation de la loi de F_n par une loi normale est correcte dès que $np \geq 10$ et $n(1-p) \geq 10$, ou dès que $np(1-p) > 18$, ou sous d'autres conditions proches, d'autant plus que n est grand et p proche de 0.5.

Lorsque p n'est pas connu, on vérifie ces conditions sur la fréquence f_n observée.

3. Intervalle de fluctuation et intervalle de confiance pour une proportion

Considérons une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$, où p est la proportion d'individus de la population ayant une propriété donnée, un échantillon $(X_1, X_2, \ldots, X_n)$ de taille n de X et la proportion (ou fréquence) d'échantillon $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre d'individus de l'échantillonnage ayant la propriété. On sait que si $np \geq 10$ et $n(1-p) \geq 10$, alors $U = \dfrac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$. On détermine comme le réel u_α tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$ grâce à la table 2. Pour $\alpha = 5\%$, on a $u_\alpha = 1.96$.

Remarque. Lorsque n est petit, on doit utiliser la loi exacte de nF_n, à savoir la loi Binomiale $\mathcal{B}(n,p)$.

3.1. Intervalle de fluctuation de la fréquence F_n

On suppose que l'on connait p.

On en déduit que $P\left(p - \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \leq F_n \leq p + \sqrt{\frac{p(1-p)}{n}}\, u_\alpha\right) = 1-\alpha$, et donc $P(F_n \in IF_p) = 1-\alpha$, avec $IF_p = \left[p - \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \,;\, p + \sqrt{\frac{p(1-p)}{n}}\, u_\alpha\right]$ **intervalle de fluctuation** IF_p **de** F_n **au niveau** $1-\alpha = 0.95$.

3.2. Intervalle de confiance de la proportion p

On suppose que l'on ne connait pas p mais que l'on a une observation f_n de F_n à partir d'un échantillon.

On a $P\left(F_n - \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \leq p \leq F_n + \sqrt{\frac{p(1-p)}{n}}\, u_\alpha\right) = 1-\alpha$, et donc $P(p \in IC_p) = 1-\alpha$, avec $IC_p = \left[F_n - \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \,;\, F_n + \sqrt{\frac{p(1-p)}{n}}\, u_\alpha\right]$ **intervalle de confiance** IC_p **de** p **au niveau** $1-\alpha = 0.95$.

Comme $\frac{F_n(1-F_n)}{n-1}$ est un estimateur sans biais de $\frac{p(1-p)}{n}$, on en déduit, si $nf_n \geq 10$ et $n(1-f_n) \geq 10$, un intervalle de confiance de la proportion p au niveau $1-\alpha$:

$$ic_p = \left[f_n - \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha \;,\; f_n + \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha\right].$$

Exemple d'intervalle de confiance

Dans une certaine espèce de rongeur, on a compté 206 mâles sur 400 naissances.

On peut considérer la situation suivante.

Population : les rongeurs d'une certaine espèce.

Variable : le sexe, à deux modalités (mâle et femelle), représenté par une variable aléatoire de loi de Bernoulli $\mathcal{B}(p)$, où p est la proportion de mâles dans la population ; on a ainsi $P(X=1) = p$ et $P(X=0) = 1-p$.

Echantillon $(X_1, X_2, \ldots, X_n)$ de taille $n = 400$ de X.

Observation de l'échantillon : $(x_1, x_2, \ldots, x_n) = (1, 1, 0, 1, \ldots, 0)$.

Estimateur de la proportion p : $F_n = \frac{\sum_{i=1}^{n} X_i}{n}$, proportion (ou fréquence) de mâles dans l'échantillon, où $\sum_{i=1}^{n} X_i$ représente le nombre de mâles de l'échantillon.

Estimation ponctuelle de la proportion p : $f_n = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{206}{400} = 0.515$, fréquence (ou proportion) de mâles dans l'observation de l'échantillon.

Intervalle de confiance de la proportion p :

$nf_n = 206 \geq 10$ et $n(1-f_n) = 194 \geq 10$

Pour $\alpha = 0,05$ (i.e. 5%), on a $u_\alpha = 1,96$.

$ic_p = \left[f_n - \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha \;;\; f_n + \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha\right] = \left[0,466 \;;\; 0,564\right]$.

Exemple d'application de l'intervalle de fluctuation

Reprenons l'exemple précédent et supposons savoir qu'il y a équiprobabilité male/femelle à chaque naissance, autrement dit que $p = 0,5$.

Pour un échantillon de $n = 400$ naissances, l'intervalle de fluctuation de F_n est $\left[p - \sqrt{\frac{p(1-p)}{n}}\, u_\alpha \,;\, p + \sqrt{\frac{p(1-p)}{n}}\, u_\alpha\right] = \left[0.5 - \sqrt{\frac{0.5(1-0.5)}{400}} \times 1.96 \,;\, 0.5 + \sqrt{\frac{0.5(1-0.5)}{400}} \times 1.96\right]$

Ainsi, 95 % des échantillons de 400 naissances donneront une fréquence d'échantillon comprise entre 0.451 et 0.551.

L'échantillon étudié donne une fréquence observée $f_n = 0.515$ qui appartient à l'intervalle de fluctuation : il est donc représentatif d'une population pour laquelle $p = 0,5$.

3.3. Intervalle de fluctuation de la fréquence F_n et loi binomiale

On considère une population dans laquelle on suppose que la proportion d'un certain caractère est p. Pour juger de cette hypothèse, on y prélève, au hasard et avec remise, un échantillon de taille n sur lequel on observe une fréquence f_n du caractère.

On rejette l'hypothèse selon laquelle la proportion dans la population est p lorsque la fréquence f_n observée est trop éloignée de p, dans un sens ou dans l'autre. On choisit de fixer le seuil de décision de sorte que la probabilité de rejeter l'hypothèse, alors qu'elle est vraie, soit inférieure à 5 %.

Lorsque la proportion dans la population vaut p, la variable aléatoire X correspondant au nombre de fois où le caractère est observé dans un échantillon aléatoire de taille n, suit la loi binomiale de paramètres n et p. On cherche à partager l'intervalle $[0,n]$, où X prend ses valeurs, en trois intervalles $[0,a-1]$, $[a,b]$ et $[b+1,n]$ de sorte que X prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur.

En tabulant les probabilités cumulées $P(X \leq k)$, pour k allant de 0 à n, il suffit de déterminer le plus petit entier a tel que $P(X \leq a) > 0,025$ et le plus petit entier b tel que $P(X \leq b) \geq 0,975$, c'est-à-dire $P(X > b) \leq 0,025$. Autrement dit, a est le plus grand entier tel que $P(X < a) \leq 0.25$. On observe aussi que $a < b$.

On a ainsi $P((X < a) \cup (X > b)) = P(X < a) + P(X > b) \leq 0.05$

et donc $P(a \leq X \leq b) = P\left(\overline{(X < a) \cup (X > b)}\right) \geq 0.95$, en étant "assez proche" de 0.95.

Comme $F_n = \frac{X}{n}$, on a ainsi $P\left(\frac{a}{n} \leq F_n \leq \frac{b}{n}\right) \geq 0.95$, en étant "assez proche" de 0.95.

La règle de décision est la suivante : si la fréquence observée f_n appartient à l'intervalle de fluctuation à 95 % $\left[\frac{a}{n}, \frac{b}{n}\right]$, on considère que l'hypothèse selon laquelle la proportion est p dans la population n'est pas remise en question et on l'accepte ; sinon, on rejette l'hypothèse selon laquelle cette proportion vaut p.

Pour $n \geq 30$, $n \times p \geq 5$ et $n \times (1-p) \geq 5$, on observe que l'intervalle de fluctuation $\left[\frac{a}{n}, \frac{b}{n}\right]$ est sensiblement le même que l'intervalle $\left[p - \frac{1}{\sqrt{n}}\ ,\ p + \frac{1}{\sqrt{n}}\right]$ proposé dans le programme de seconde.

Exemple d'exercice

Monsieur Z, chef du gouvernement d'un pays lointain, affirme que 52 % des électeurs lui font confiance. On interroge 100 électeurs au hasard (la population est suffisamment grande pour considérer qu'il s'agit de tirages avec remise) et on souhaite savoir à partir de quelles fréquences, au seuil de 5 %, on peut mettre en doute le pourcentage annoncé par Monsieur Z, dans un sens, ou dans l'autre.

1. On fait l'hypothèse que Monsieur Z dit vrai et que la proportion des électeurs qui lui font confiance dans la population est $p = 0,52$. Montrer que la variable aléatoire X, correspondant au nombre d'électeurs lui faisant confiance dans un échantillon de 100 électeurs, suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

2. On donne ci-contre un extrait de la table des probabilités cumulées $P(X \leq k)$ où X suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

Déterminer a et b tels que définis précédemment et comparer les intervalles de fluctuation à 95 % $\left[\frac{a}{n}, \frac{b}{n}\right]$ et $\left[p - \frac{1}{\sqrt{n}}\ ,\ p + \frac{1}{\sqrt{n}}\right]$.

k	$P(X \leq k)$
40	0,0106
41	0,0177
42	0,0286
43	0,0444
...	
61	0,9719
62	0,9827
63	0,9897
64	0,9941

3. Énoncer la règle décision permettant de rejeter ou non l'hypothèse $p = 0,52$, selon la valeur de la fréquence f des électeurs favorables à Monsieur Z obtenue sur l'échantillon.

4. Sur les 100 électeurs interrogés au hasard, 43 déclarent avoir confiance en Monsieur Z. Peut-on considérer, au seuil de 5 %, l'affirmation de Monsieur Z comme exacte ?

Remarque : la recherche de l'intervalle de fluctuation peut-être illustrée par le diagramme en bâton de la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

Utilisation du tableur Excel

Construire la table des probabilités et des probabilités cumulées de la loi Binomiale de paramètres $n = 100$ et $p = 0,52$. Construire le diagramme en bâton de cette loi.

4. Test de conformité pour une proportion p

On s'intéresse à la question suivante : étant donnée une population dans laquelle une proportion p d'individu ont une certaine propriéte, peut-on raisonnablement supposer que p est égal à une certaine valeur p_0 donnée a priori ?

Par exemple, des tests en laboratoire permettent d'affirmer qu'un certain médicament est efficace sur une proportion p_0 d'individus atteints d'une certaine maladie. Mais après sa mise sur le marché, le médicament a-t-il la même efficacité sur l'ensemble des individus malades ? Comment savoir si la proportion p de malades guéris par le médicament est bien égale à p_0 ?

La réponse à la question est donnée par la mise en place d'un test de conformité.

De façon générale, un **test statistique** est une procédure permettant de calculer la valeur d'une certaine fonction des observations d'un ou de plusieurs échantillon, qui conduit à rejeter ou non, avec un certain risque d'erreur, une hypothèse généralement appelée **hypothèse nulle** et notée H_0. Celle-ci porte sur la (ou les) population(s) d'où est (sont) issu(s) le(s) échantillon(s). Elle s'oppose à une **hypothèse** dite **alternative** et notée H_1.

Considérons une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$, où p est la proportion d'individus de la population ayant une propriété donnée, un échantillon $(X_1, X_2, \dots, X_n)$ de taille n de X et la proportion (ou fréquence) d'échantillon $F = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre d'individus de l'échantillonnage ayant la propriété. On sait que si $np \geq 10$ et $n(1-p) \geq 10$, alors $U = \dfrac{F-p}{\sqrt{\frac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$.

Test (bilatéral) **de** $H_0 : p = p_0$ **contre** $H_1 : p \neq p_0$.

On utilise alors une variable aléatoire dont on connait la loi de probabilité lorsque H_0 est vraie. Par exemple $U = \dfrac{F-p}{\sqrt{\frac{p(1-p)}{n}}}$, car lorsque H_0 est vraie, on sait que $U = \dfrac{F-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ suit la loi $\mathcal{N}(0;1)$.

On fixe une valeur $\alpha \in]0,1[$. En général, on prend α petit, le plus souvent $0,05$, $0,01$, $0,001$. On peut trouver un réel u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$. Ce réel u_α peut être trouvé dans la table 2.

On est donc amené à comparer la proportion F de l'échantillon à la proportion théorique $p = p_0$. L'hypothèse H_0 signifiera que les différences observées sont seulement dûes aux fluctuations d'échantillonnage (i.e. ne sont pas significatives).

On ne rejettera pas H_0 si les différences observées ne sont pas significatives, c'est-à-dire si U est "petite", ce que l'on peut traduire par $-u_\alpha < U < u_\alpha$, c'est-à-dire $|U| < u_\alpha$.

On rejetera donc H_0 si les différences observées sont significatives, ce que l'on peut traduire par $U > u_\alpha$ ou $U < -u_\alpha$, c'est-à-dire $|U| > u_\alpha$. Par construction de u_α, on a $P(U > u_\alpha) = P(U < -u_\alpha) = \dfrac{\alpha}{2}$, soit encore $P(|U| > u_\alpha) = \alpha$, i.e. $P(U \notin]-u_\alpha, u_\alpha[) = \alpha$.

En pratique, on calcule $u = \dfrac{f-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ et on décide

- de rejeter H_0 si $u \notin]-u_\alpha, u_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-u_\alpha, u_\alpha[$ aurait une probabilité faible de se réaliser ; on pourra dire que la valeur observée $\bar{x}$ n'est pas conforme à la valeur théorique μ_0 mais on ne pourra pas donner de valeur acceptable de μ ;
- de ne pas rejeter H_0 si $u \in]-u_\alpha, u_\alpha[$, car si H_0 était vraie, l'événement $U \in]-u_\alpha, u_\alpha[$ aurait une probabilité forte de se réaliser ; on pourra dire que la valeur observée f est conforme à la valeur théorique p_0 et que la valeur p_0 ne peut être rejeter. Attention : d'autres valeurs p_0', p_0'', ... peuvent également convenir.

Erreurs de décision.

Lorsqu'on rejette H_0 alors que H_0 est vraie, on commet une erreur. On a donc une probabilité α de se tromper : α est appelée **erreur de première espèce**. En effet, lorsque H_0 est vraie, on a $P(U \notin]-u_\alpha, u_\alpha[) = \alpha$.

Lorsque l'on ne rejette pas H_0 alors que H_0 est fausse, on commet une erreur. On a une probabilité β de se tromper : β est appelée **erreur de deuxième espèce**. Cette erreur est difficilement calculable. La plupart du temps, on ne connait pas la loi de U lorsque H_0 est fausse. La valeur $1-\beta$ est appelée la **puissance du test**.

Test (bilatéral) **de** $H_0 : p = p_0$ **contre** $H_1 : p \neq p_0$.
On calcule $u = \dfrac{f - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$. On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, et on décide que :
- si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Exemple
Reprenons l'exemple précédents sur les rongeurs.
Sur un échantillon de 400 naissances, on a observé 206 mâles, soit une fréquence de mâles de $f = \dfrac{206}{400} = 0.515$.
On se demande s'il y a autant de mâles que de femelles dans la population ; autrement dit si la proportion de mâles dans la population est $p = 0.5$.
On peut effectuer le test statistique de $H_0 : p = p_0$ **contre** $H_1 : p \neq p_0$, **avec** $p_0 = 0.5$.
On calcule $u = \dfrac{f - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \dfrac{0.515 - 0.5}{\sqrt{\dfrac{0.5(1-0.5)}{400}}} = 0.6$. Pour $\alpha = 0,05$ (i.e. 5%), on a $u_\alpha = 1,96$.
Comme $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 : il est donc possible que $p = 0.5$.

Test (unilatéral) **de** $H_0 : p = p_0$ **contre** $H_1 : p > p_0$.
On détermine u'_α tel que $P(U < u'_\alpha) = 1-\alpha$, i.e. $u'_\alpha = \phi^{-1}(1-\alpha) = u_{2\alpha}$, et on décide que :
- si $u < u'_\alpha$, alors on ne peut rejeter H_0 ;
- si $u \geq u'_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) **de** $H_0 : p = p_0$ **contre** $H_1 : p < p_0$.
On détermine u''_α tel que $P(U \geq u''_\alpha) = 1-\alpha$, i.e. $u''_\alpha = \phi^{-1}(\alpha) = u_{2-2\alpha} = -u_{2\alpha}$, et on décide que :
- si $u > u''_\alpha$, alors on ne peut rejeter H_0 ;
- si $u \leq u''_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Remarque : **la** p**-valeur**.
L'utilisation de logiciels (tels R) évite d'avoir à mener ces calculs. Lors de la mise en oeuvre de tout test avec un logiciel (cela sera valable dans tous les chapitres suivants de ce cours), ce dernier fournit souvent la p-**valeur** qui sera comparée au risque α pour la prise de décision. Ainsi, lorsque $p \leq \alpha$, on rejette H_0 au risque α ; lorsque $p > \alpha$, on ne peut rejeter H_0 au risque α.
L'interprétation de la p-valeur est simple : plus elle est faible, plus la décision de rejeter H_0 est fiable.

5. Comparaison de deux proportions

Dans deux populations P_1 et P_2 on étudie le même caractère "avoir ou non une propriété donnée". Soient X_1 et X_2 des variables aléatoires de loi de Bernoulli $\mathcal{B}(p_1)$ et $\mathcal{B}(p_2)$ représentant le caractère dans chaque population, où p_1 (respectivement p_2) est la proportion d'individus ayant la propriété dans P_1 (respectivement dans P_2). De P_1 et P_2 on extrait un échantillon $E_1 = (X_{1,1}, X_{1,2}, \ldots, X_{1,n_1})$ de taille n_1 de X_1 et un échantillon $E_2 = (X_{2,1}, X_{2,2}, \ldots, X_{2,n_2})$ de taille n_2 de X_2.

Les fréquences d'échantillon sont alors $F_1 = \dfrac{\sum_{i=1}^{n_1} X_{1,i}}{n_1}$ et $F_2 = \dfrac{\sum_{i=1}^{n_2} X_{2,i}}{n_2}$.

5.1. Cas d'échantillons indépendants

Les échantillons E_1 et E_2 sont supposés indépendants.

Test (bilatéral) **de** $H_0 : p_1 = p_2 = p$ **contre** $H_1 : p_1 \neq p_2$.

Supposons que $n_1 f_1 \geq 5$, $n_1(1-f_1) \geq 5$, $n_2 f_2 \geq 5$, $n_2(1-f_2) \geq 5$. Sous l'hypothèse H_0, $U = \dfrac{F_1 - F_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$, et en regroupant les deux échantillons, on peut estimer p par $f_{1,2} = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$. On calcule $u = \dfrac{f_1 - f_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})f_{1,2}(1-f_{1,2})}}$. On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$, i.e. $u_\alpha = \phi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$ (table 2) et on décide que :

- si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) **de** $H_0 : p_1 = p_2$ **contre** $H_1 : p_1 > p_2$.

On détermine u'_α tel que $P(U < u'_\alpha) = 1 - \alpha$, i.e. $u'_\alpha = \phi^{-1}(1-\alpha) = u_{2\alpha}$, et on décide que :

- si $u < u'_\alpha$, alors on ne peut rejeter H_0 ;
- si $u \geq u'_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) **de** $H_0 : p_1 = p_2$ **contre** $H_1 : p_1 < p_2$.

On détermine u''_α tel que $P(U \geq u''_\alpha) = 1 - \alpha$, i.e. $u''_\alpha = \phi^{-1}(\alpha) = u_{2-2\alpha} = -u_{2\alpha}$, et on décide que :

- si $u > u''_\alpha$, alors on ne peut rejeter H_0 ;
- si $u \leq u''_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

5.2. Cas d'échantillons appariés : test de McNemar

Deux échantillons E_1 et E_2 sont dits *appariés* lorsque chaque observation $x_{1,i}$ de E_1 est associée à une valeur $x_{2,i}$ de E_2 (appariés = associés par paires). C'est par exemple le cas lorsque E_1 et E_2 proviennent d'un même groupe de malades avant et après traitement. Deux échantillons appariés ont donc la même taille $n_1 = n_2 = n$.

On utilise le tableau suivant des effectifs de présence ou absence de la propriété étudiée :

$P_1 \setminus P_2$	présent	absent	totaux
présent	a	b	$a+b$
absent	c	d	$c+d$
totaux	$a+c$	$b+d$	n

Le test de McNemar s'appuie sur le calcul de $u = \dfrac{b-c}{\sqrt{b+c}}$, et se poursuit de façon analogue au cas d'échantillons indépendants (paragraphe 5.1). On peut l'utiliser dès que $b + c \geq 10$.

5.3. Exemple

Dans une même catégorie sociale, un échantillon de 40 hommes a fourni 8 fumeurs et un échantillon de 60 femmes a fourni 18 fumeuses. On se demande si la proportion de fumeurs est la même pour les deux sexes.

On peut considérer la situation suivante.

Population 1 : hommes.

Variable X_1 : être fumeur, représenté par une variable aléatoire X_1 de loi de Bernoulli $\mathcal{B}(p_1)$, où p_1 est la proportion d'hommes fumeurs.

Echantillon de taille $n_1 = 40$.

Estimateur de p_1 : fréquence d'échantillon F_1. Estimation de p_1 : $f_1 = \frac{8}{40} = 0,2$.

Population 2 : femmes.

Variable X_2 : être fumeuse, représenté par une variable aléatoire X_2 de loi de Bernoulli $\mathcal{B}(p_2)$, où p_2 est la proportion de femmes fumeuses

Echantillon de taille $n_2 = 60$.

Estimateur de p_2 : fréquence d'échantillon F_2. Estimation de p_2 : $f_2 = \frac{18}{60} = 0,3$.

Les échantillons E_1 et E_2 sont indépendants.

Test (bilatéral) de $H_0 : p_1 = p_2 = p$ contre $H_1 : p_1 \neq p_2$.

Supposons que $n_1 f_1 = 8 \geq 5$, $n_1(1-f_1) = 32 \geq 5$, $n_2 f_2 = 18 \geq 5$, $n_2(1-f_2) = 42 \geq 5$.

Sous l'hypothèse H_0, $U = \frac{F_1 - F_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})p(1-p)}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$, et en regroupant les deux échantillons, on peut estimer p par $f_{1,2} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{8+18}{40+60} = 0,26$. En remplaçant p par $f_{1,2}$, on ne modifie pas la loi approchée de U.

On calcule $u = \frac{f_1 - f_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})f_{1,2}(1-f_{1,2})}} = \frac{0,2-0,3}{\sqrt{(\frac{1}{40} + \frac{1}{60})0,26(1-0,26)}} \simeq -1,12$.

On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$ (table 2) : pour $\alpha = 0,05$, on trouve $u_\alpha = 1,96$.

Comme $u \in]-u_\alpha, u_\alpha[$, on ne peut rejeter H_0 : la proportion de fumeurs ne diffère pas significativement entre les deux sexes. Pour cette décision de non-rejet, on ne connait pas la probabilité de se tromper (erreur de deuxième espèce).

6. Exercices

Exercice 1.

On admet que dans la population d'enfants de 11 à 14 ans d'un département français, le pourcentage d'enfants ayant déjà eu une crise d'asthme dans leur vie est de 13%.

Un médecin d'une ville de ce département est surpris par le nombre important d'enfants le consultant pour des crises d'asthmes. Il décide de mener une étude statistique en choisissant de manière aléatoire 100 enfants de 11 à 14 ans de la ville. Il observe que 19 d'entre eux ont déjà eu une crise d'asthmes.

1) Utiliser un intervalle de fluctuation pour aider le médecin à décider s'il y a plus d'enfants ayant des crises d'asthmes dans la ville que dans le département.

2) Le médecin n'est pas convaincu par la décision obtenue et pense que le nombre d'enfants interrogés était insuffisant. Combien d'enfants faudrait-il interroger pour qu'une fréquence observée de 0,19 amène à conclure qu'il y a plus d'enfants ayant des crises d'asthmes dans la ville que dans le département.

Exercice 2.

Un groupe d'étudiants en Statistique réalise une enquête auprès d'une population d'étudiants en sociologie en interrogeant un échantillon de 135 individus. Ils désirent connaître, entre autres, la proportion p d'étudiants ayant suivi des études secondaires scientifiques.

Pour accélérer le traitement, ils partagent le dépouillement en deux groupes. Un groupe constate que sur 60 des étudiants interrogés, 24 ont suivi des études secondaires scientifiques. L'autre groupe constate que sur les 75 des étudiants interrogés restant, 33 ont suivi des études secondaires scientifiques.

1) Déterminer trois estimations ponctuelles de p.

2) A partir de l'échantillon des 135 étudiants, déterminer un intervalle de confiance de p au seuil $\alpha = 5\%$.

3) On souhaite estimer p avec une précision de $0,05$. Quelle devrait être la taille n de l'échantillon ?

Exercice 3.

Pour obtenir une estimation de la proportion d'hyperglycémiques parmi les personnes âgées de plus de soixante ans (population P), on choisit au hasard 170 personnes dans P. On constate que parmi celles-ci, 31 sont hyperglycémiques.

1) Donnez un intervalle de confiance au niveau 95% pour la proportion p de personnes hyperglycémiques de P.

2) Si on effectuait 200 fois le tirage de 170 personnes de P, on pourrait construire 200 intervalles de confiance du type précédent. Parmi ces 200 intervalles, combien, en moyenne, contiendraient la valeur de p ?

Exercice 4.

Un sondage effectué sur un échantillon de 400 électeurs donne 212 intentions de vote en faveur d'un candidat C.

1) Déterminer un intervalle de confiance au niveau 95% pour la proportion d'électeurs, dans l'ensemble de la population électorale, ayant l'intention de voter en faveur de C.

2) Quelle taille minimale de l'échantillon faudrait-il prendre pour que l'intervalle (au même niveau 95%) ne contienne pas la valeur 0,50 ?

Exercice 5.

Lors d'une précédente consultation électorale, le candidat *A* avait obtenu 51% des suffrages exprimés. A l'approche de nouvelles élections, il réalise un sondage sur un échantillon de 400 électeurs choisis au hasard dans sa circonscription. Il obtient 196 intentions de votes.

Peut-il conclure que sa cote de popularité est restée stable ?

Exercice 6.

Une agence de publicité affirme qu'un produit d'entretien est efficace à 90% pour déboucher éviers et lavabos en deux heures, quelle que soit la nature de l'obstruction. Une association de défense du consommateur a fait une enquête qui relève que sur 100 lavabos bouchés, 80 seulement sont débouchés en deux heures en utilisant le produit d'entretien.

1) Doit-on faire un procès à l'agence de publicité ? Faire un test au risque 5%, puis 1%.
2) En utilisant le logiciel R, on a obtenu les résultats suivants :

```
> prop.test(80,100,0.9,alternative="less")

        1-sample proportions test with continuity correction

data:  80 out of 100, null probability 0.9
X-squared = 10.0278, df = 1, p-value = 0.000771
alternative hypothesis: true p is less than 0.9
95 percent confidence interval:
 0.0000000 0.8617706
sample estimates:
  p
0.8
```

Cela confirme-t-il les résultats du 1) ?

Exercice 7.

On compare les effets d'un même traitement dans deux hopitaux différents. Dans le premier hopital, 70 des 100 malades traités montrent des signes de guérison. Dans le deuxième hopital, c'est le cas pour 100 des 150 malades traités.

Quelle conclusion peut-on en tirer ?

Exercice 8. *D'après examen de mars 2011*

Afin d'évaluer l'impact d'une campagne média anti-tabac, on s'est intéressé à la proportion de fumeurs menant des actions pour essayer d'arrêter de fumer (diminution de la consommation, achat de patchs anti-tabac, consultations médicales, ...), c'est-à-dire à la proportion de fumeurs "actifs" pour arrêter.

Un sondage "avant campagne" a été effectué auprès de 3000 fumeurs, et un sondage "après campagne" a été effectué auprès d'un autre échantillon de 3000 fumeurs ; les deux échantillons sont donc indépendants.

Le premier sondage donne une proportion de 0,15 de fumeurs "actifs", alors que le deuxième sondage donne une proportion de 0,17 de fumeurs "actifs".

On veut savoir si la campagne a été efficace ; autrement dit si la proportion de fumeurs "actifs" a augmenté après la campagne.

1) a) Déterminer un intervalle de confiance au niveau 95% de la proportion de fumeurs "actifs" avant la campagne. Préciser la population et le caractère étudié, la taille d'échantillon, le(s) estimateur(s) mis en jeu.

b) De façon analogue, donner (sans détailler les calculs) un intervalle de confiance au niveau 95% de la proportion de fumeurs "actifs" après la campagne.

c) Peut-on déduire de ces deux intervalles que la campagne a été efficace ?

2) a) Expliquer brièvement ce que représentent les erreurs de première et deuxième espèce d'un test statistique.

b) Effectuer un test statistique au risque 5%, puis 10%, pour savoir si la campagne a été efficace. En cas de décisions contradictoires avec les deux risques 5% et 10%, préciser et justifier la décision à retenir.

Exercice 9.

Sous forme de comprimé un médicament est efficace dans le traitement d'une maladie dans 80% des cas. Le pharmacien du laboratoire qui commercialise ce médicament, essaie une forme injectable par voie intra-musculaire, de ce même médicament. Il observe sur un échantillon de 50 malades, 35 guérisons. L'efficacité de la forme intra-musculaire est-elle différente de celle en comprimé ? Lui est-elle inférieure ? (conclure au risque de 5%).

Exercice 10.

On sait qu'une maladie atteint 10% des individus d'une population *P* donnée. Un chercheur a expérimenté un traitement sur un échantillon de *n* individus : il a alors recensé 5% de malades. Déterminer la valeur minimale de *n* qui permette au chercheur de conclure à l'efficacité du traitement au risque de 5%.

Exercice 11.

Pour traiter un certain type de tumeur, on a utilisé deux schémas thérapeutiques :

- sur 40 malades traités avec le schéma A, on a observé une mortalité à 5 ans de 15 % ;
- sur 60 malades traités avec le schéma B, on a observé une mortalité à 5 ans de 25 %.

Si l'on considère la mortalité à 5 ans, peut-on dire que les schémas A et B diffèrent significativement au risque 10 % ? au risque 5 % ?

Exercice 12.

Pour comparer deux somnifères A et B, on procède aux deux expériences suivantes.

1) Dans la première expérience, on compare deux séries, constituées par tirage au sort, de 64 sujets ayant reçu A pour la première, de 64 sujets ayant reçu B pour la seconde. Considérant qu'un soporifique donne un succès en cas de durée de sommeil supérieure ou égale à 6 heures, on obtient les résultats suivants :

	A	B
Succès	42	32
Echec	22	32

Au risque 5%, les différences observées entre les deux somnifères sont-elles significatives ?

2) Dans la deuxème expérience, on utilise 64 sujets recevant une fois A et une fois B, dans un ordre tiré au sort. On obtient les résultats suivants :

	Succès avec B	Echec avec B
Succès avec A	27	15
Echec avec A	5	17

Les somnifères A et B ont-ils la même efficacité ?

TABLE 1

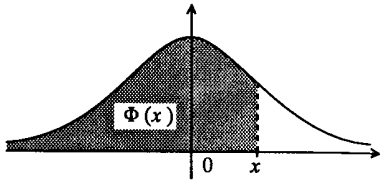
Fonction de répartition de la loi normale réduite

Si U suit la loi normale réduite, pour $x \geqslant 0$, la table donne la valeur :

$\phi(x) = P(U \leqslant x)$.

La valeur x s'obtient par addition des nombres inscrits en marge.
Pour $x < 0$, on a :

$\phi(x) = 1 - \phi(-x)$.



x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 1	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 9	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7
2,0	0,977 2	0,977 8	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 5	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6

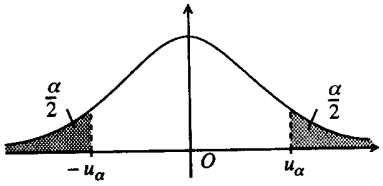
TABLE 2

Loi normale réduite (table de l'écart réduit)

Si U est une variable aléatoire qui suit la loi normale réduite, la table donne, pour α choisi, la valeur u_α telle que :

$P(|U| \geqslant u_\alpha) = \alpha$.

La valeur α s'obtient par addition des nombres inscrits en marge.



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013