Tests d'hypothèses, Intervalle de confiance

Exercice 1. Un échantillon de 10 000 personnes sur une population étant donné, on sait que le taux moyen de personnes à soigner pour un problème de cholestérol élevé est de $7,5\%$. Donner un intervalle dans lequel on soit «sûr» à 95%, de trouver le nombre exact de personnes à soigner sur les 10 000.

Correction 1. Un intervalle dans lequel on soit «sûr» à 95% de trouver le nombre exact de personnes à soigner sur les 10 000 : $[p - y_\alpha\sqrt{\frac{p(1-p)}{n}}; fe + y_\alpha\sqrt{\frac{p(1-p)}{n}}]$. Fréquence entre 65,7% et 94,3%. Donc entre 698 et 802 personnes sur 10000

Exercice 2. Un vol Marseille - Paris est assuré par un Airbus de 150 places ; pour ce vol des estimations ont montré que la probabilité pour qu'une personne confirme son billet est $p = 0.75$. La compagnie vend n billets, $n > 150$. Soit X la variable aléatoire «nombre de personnes parmi les n possibles, ayant confirmé leur réservation pour ce vol».

1. Quelle est la loi exacte suivie par X ?
2. Quel est le nombre maximum de places que la compagnie peut vendre pour que, à au moins 95%, elle soit sûre que tout le monde puisse monter dans l'avion, c'est-à-dire n tel que : $P[X > 150] \leq 0.05$?
3. Reprendre le même exercice avec un avion de capacité de 300 places ; faites varier le paramètre $p = 0.5$; $p = 0.8$.

Correction 2. La loi exacte suivie par X est une loi binomiale de paramètres : n, p. $E(X) = 0.75n$ et $\operatorname{Var} X = 0.25 \cdot 0.75n$. Comme $n > 150$, on peut faire l'approximation par la loi normale d'espérance $0,75n$ et d'écart-type $\sigma = \sqrt{0.25 \cdot 0.75n}$. $P[X > 150] \leq 0.05$ si $P[X \leq 150] \geq 0.95$ si : $P[\frac{X-0.75n}{\sqrt{0.25 \cdot 0.75n}} \leq \frac{150-0.75n}{\sqrt{0.25 \cdot 0.75n}}] \geq 0.95$. Dans la table de Gauss, on lit $F(1.645) = 0.95$. On n'a plus qu'à résoudre l'inéquation : $\frac{150.5-0.75n}{\sqrt{0.25 \cdot 0.75n}} \geq 1.645$, dont les solutions sont :

$$0 \leq n \leq 187.$$

Ainsi, en vendant moins de 187 billets, la compagnie ne prend qu'un risque inférieur à 5% de devoir indemniser des voyageurs en surnombre. Faisons varier les paramètres, cela ne pose aucun problème :
$N = 150$, $p = 0.5$. n est solution de l'inéquation : $\frac{150.5-0.5n}{\sqrt{0.5.0.5n}} \geq 1.645$. Solution : $n \leq 272$.
$N = 300$, $p = 0.75$. n est solution de l'inéquation : $\frac{300.5-0.75n}{\sqrt{0.25.0.75n}} \geq 1.645$. Solution : $n \leq 381$.
$N = 300$, $p = 0.5$. n est solution de l'inéquation : $\frac{300.5-0.5n}{\sqrt{0.5.0.5n}} \geq 1.645$. Solution : $n \leq 561$.

Exercice 3. Un petit avion (liaison Saint Brieuc-Jersey) peut accueillir chaque jour 30 personnes ; des statistiques montrent que 20% des clients ayant réservé ne viennent pas. Soit X la variable aléatoire : «nombre de clients qui se présentent au comptoir parmi 30 personnes qui ont réservé».

1. Quelle est la loi de X ? (on ne donnera que la forme générale) ; quelle est son espérance, son écart-type ?
2. Donner un intervalle de confiance au seuil 95%, permettant d'estimer le nombre de clients à prévoir.

Correction 3.
1. La loi de X est la loi binomiale $n = 30$, $p = 0.2$.
2. Un intervalle de confiance au seuil 95%, permettant d'estimer le nombre de clients à prévoir : c'est pour la fréquence : 0.657 ; 0.943. Soit entre 20 et 28 personnes. C'est une large fouchette due à n petit.

Exercice 4. Une compagnie aérienne a demandé des statistiques afin d'améliorer la sûreté au décollage et définir un poids limite de bagages. Pour l'estimation du poids des voyageurs et du poids des bagages, un échantillon est constitué de 300 passagers qui ont accepté d'être pesés : on a obtenu une moyenne m_e de 68kg, avec un écart-type σ_e de 7 kg.

1. Définir un intervalle de confiance pour la moyenne des passagers. (On admet que le poids des passagers suit une loi normale de moyenne m, d'écart-type σ.)
2. Montrer que l'on peut considérer que le poids des passagers est une variable aléatoire X de moyenne 70 kg, d'écart-type 8 kg.
3. En procédant de même pour le poids des bagages, on admet les résultats :
   - Si le poids maximum autorisé est de 20 kg, le poids des bagages peut être considéré comme une variable aléatoire Y de moyenne 15 kg, d'écart-type 5 kg.
   - La capacité de l'avion est de 300 passagers ; l'avion pèse, à vide, 250 tonnes. Le décollage est interdit si le poids total dépasse 276.2 tonnes. Quelle est la probabilité pour que le décollage soit interdit ?

Correction 4.
1. On peut estimer m par la moyenne de l'échantillon : 68 kg, et σ par $\sigma_e\sqrt{\frac{300}{299}} = 7\sqrt{\frac{300}{299}} \simeq 7.0117$ kg. On en déduit un intervalle de confiance pour la moyenne m : $I_\alpha = [67.2; 68.8]$.
2. La borne supérieure de l'intervalle étant de 69 kg, il est raisonnable de prendre 70 kg comme espérance de la variable poids d'un passager.
3. Le décollage est autorisé si le poids total des voyageurs et de leurs bagages ne dépasse pas 26.2 tonnes. Pour chacun des 300 passagers, notons : X_i son poids et Y_i le poids de ses bagages. Faisons l'hypothèse d'indépendance entre les variables X_i et Y_i. Le poids total

$Z = \sum_{i=1}^{300}(X_i + Y_i)$ est la somme de 600 variables aléatoires indépendantes ; le théorème central limite s'applique sous cette hypothèse. Comme l'espérance totale est $E(Z) = 300 \cdot (70 + 15) = 25\,500$ et la variance de Z est : $\operatorname{Var} Z = 300 \cdot (\operatorname{Var} X_i + \operatorname{Var} Y_i)$. Alors Z suit approximativement une loi normale de moyenne $m = 25\,500$, d'écart-type $\sigma = \sqrt{300 \cdot (8^2 + 5^2)} = 163.4$. Alors $Z' = \frac{Z-m}{\sigma}$ suit approximativement une loi normale centrée réduite. Le décollage est interdit si : $Z > 26\,200$, c'est-à-dire si $Z' > 4.284$. On lit dans la table de Gauss : pour $t = 4$, $F(t) = 0.999\,968 = P[Z' \leq 4]$. Le décollage est interdit pour cause de surcharge pondérale avec une probabilité inférieure à $0.000\,04$.

Exercice 5. Afin de mieux satisfaire leurs clients, une grande société fournisseur d'accès internet fait ses statistiques sur le nombre d'appels reçus en *hotline*, elle pourra ainsi évaluer le temps d'attente pour le client et le nombre d'employés à mettre au standard ; les résultats de l'enquête portent sur 200 séquences consécutives de une minute, durant lesquelles le nombre d'appels moyen a été de 3 appels par minute. On suppose que les appels sont répartis également dans le temps : on partage un intervalle de temps en unités de une seconde ; alors dans chaque unité de temps, il y a au plus un appel.

1. Quelle est la loi de probabilité du nombre d'appels reçus en 4 minutes ?
2. Montrer que l'on peut approcher cette loi par une loi de Poisson.
3. Donner un intervalle de confiance pour le nombre moyen d'appels en 4 minutes.

Correction 5.
1. L'intervalle de temps de 4 minutes est la répétition de 240 secondes, au cours desquelles les appels surviennent de façon indépendante, avec la probabilité d'appel de $\frac{1}{20}$; la loi de probabilité du nombre d'appels reçus en 4 minutes est donc une loi binomiale, de paramètres $n = 240$ et $p = \frac{1}{20}$.
2. Comme $n \geq 30$ et $np \leq 15$, il est possible d'approcher cette loi par une loi de Poisson de paramètre λ estimé par $np = 12$.
3. Un échantillon de taille 200 a été réalisé pour estimer le nombre moyen d'appels par minute ; c'est un échantillon de taille 50 pour la variable précédente (nombre d'appels reçus en 4 minutes) qui suit une loi de Poisson d'espérance et de variance 12. Un intervalle de confiance au niveau 95% pour la moyenne est $I_\alpha = [11; 13]$.

Exercice 6. On s'intéresse au problème des algues toxiques qui atteignent certaines plages de France ; après étude on constate que 10% des plages sont atteintes par ce type d'algues et on veut tester l'influence de rejets chimiques nouveaux sur l'apparition de ces algues. Pour cela 50 plages proches de zones de rejet chimiques, sont observées ; on compte alors le nombre de plages atteintes par l'algue nocive : on constate que 10 plages sont atteintes par

l'algue. Pouvez-vous répondre à la question «Les rejets chimiques ont-t-il modifié, de façon significative, avec le risque $\alpha = 0.05$, le nombre de plages atteintes ?»

Correction 6. Posons H_0 «les rejets chimiques ne modifient pas le nombre de plages atteintes par les algues».
Notons $p_0 = 0.1$ la proportion théorique de plages atteintes par l'algue verte avant les rejets chimiques ; p la proportion théorique de plages atteintes par l'algue verte après les rejets chimiques et f la fréquence observée dans l'échantillon.
Considérons alors la variable aléatoire X_i, $i \leq 50$, qui a deux modalités : 1 si la plage est atteinte, 0 sinon. C'est une variable de Bernoulli, alors le nombre total de plages atteintes dans l'échantillon est une variable aléatoire qui, sous H_0, obéit à une loi binomiale de paramètres $n = 50$, $p_0 = 0.1$.
Sous H_0, «$p = p_0 = 0.1$» la variable «moyenne d'échantillon» :

$$\overline{X} = \frac{\sum_{i=1}^{i=50} X_i}{n}$$

dont une réalisation est la fréquence observée, soit $\frac{10}{50}$, obéit à une loi que l'on peut approcher par une loi normale de paramètres : moyenne p_0 et écart-type $\sqrt{\frac{p_0(1-p_0)}{50}}$.
A l'aide de la formule de cours, on détermine l'intervalle de confiance associé : $I \simeq [0.017; 0.183]$. On constate que la fréquence observée est dans la zone de rejet (non chimique) : 0.2 n'est pas dans l'intervalle de confiance au seuil 95%. On peut donc rejeter H_0 et conclure, au risque 0.05, que les rejets chimiques modifient de façon significative le nombre de plages atteintes par l'algue.

Exercice 7. On veut étudier la liaison entre les caractères : «être fumeur» (plus de 20 cigarettes par jour, pendant 10 ans) et «avoir un cancer de la gorge», sur une population de 1000 personnes, dont 500 sont atteintes d'un cancer de la gorge. Voici les résultats observés :
Tableau observé

Observé	cancer	non cancer	marge
fumeur	342	258	600
non fumeur	158	242	400
marge	500	500	1000

Faire un test d'indépendance pour établir la liaison entre ces caractères.

Correction 7. Mise en oeuvre du test :

1. On définit un risque : 5%. Pour étudier la dépendance de ces caractères faisons l'hypothèse H_0 : «les deux caractères sont indépendants » et voyons ce qui se passerait sous cette hypothèse. Notons les événements :

– C : «avoir un cancer dans la population observée»
– F : «être fumeur dans la population observée»
Si les événements F et C sont indépendants, alors : $P(F \cap C) = P(F) \cdot P(C)$ et de même pour les trois autres possibilités : $P(\overline{C} \cap F)$, $P(\overline{C} \cap \overline{F}), P(C \cap \overline{F})$, quantités que l'on peut donc calculer sous H_0 :
$P(F) = \frac{600}{1000}$, $P(C) = \frac{500}{1000}$, $P(F) \cdot P(C) = \frac{3}{10}$, alors l'effectif théorique correspondant à la catégorie «fumeur et cancéreux» est de 300.

2. On en déduit le tableau théorique sous H_0 :

Théorique	cancer	non cancer	marge
fumeur	300	300	600
non fumeur	200	200	400
marge	500	500	1000

3. On calcule alors la valeur de $s = \sum_{i=1}^{i=4} \frac{(O_i - T_i)^2}{T_i}$: on obtient : $s = 34.73$. On a précisé le risque de %, mais pour $\alpha = 0,001$, on lit dans la table du khi-deux à un degré de liberté : $P[\chi^2 \geq 10.83] = 0.001$ et le χ^2 calculé est 34.73 !

4. On décide de rejeter H_0. Ainsi, en rejetant l'hypothèse de l'indépendance des caractères «être fumeur» et «avoir un cancer de la gorge», on a moins de une chance sur 1000 de se tromper, puisque moins de un tableau possible sur mille conduit à un calcul de χ^2 plus grand que 10.83 ; beaucoup moins sans doute, conduiraient à un calcul de χ^2 plus grand que 34.73.