

Feuille 4
Statistique : estimation et tests

Exercice 1.

On joue avec un dé cubique parfait. À partir de combien de lancers peut-on affirmer, avec un risque de se tromper inférieur à 5%, que la fréquence d'apparition de 1, au cours de ces lancers, ne différera pas de $\frac{1}{6}$ de plus de $\frac{1}{100}$?

Exercice 2.

Une urne contient N boules, numérotées de 1 à N . Le but de l'exercice est d'estimer, à partir de tirages effectués dans cette urne, le nombre total de boules N . On effectue donc n tirages indépendants avec remise, et pour tout $i \in \{1, \dots, n\}$, on note X_i la variable aléatoire associée au numéro obtenu au i -ème tirage, et on définit $\hat{N}_n^{(\max)} = \max(X_1, \dots, X_n)$.

1. Déterminer la loi de X_i .
2. Calculer la fonction de répartition $F_{\hat{N}_n^{(\max)}}$ de $\hat{N}_n^{(\max)}$.
3. Calculer la probabilité que $\hat{N}_n^{(\max)}$ soit égal à N . Quelle est la limite de cette probabilité quand le nombre de tirages n tend vers l'infini ?
4. En utilisant $\hat{N}_n^{(\max)}$ comme statistique de test, construire un test de niveau $\alpha = 0.15$ de l'hypothèse $H_0 : \ll N = 99 \gg$ contre $H_1 : \ll N = 101 \gg$.
5. On a observé les numéros suivants :

12	52	48
25	64	87

Quelle est dans ce cas la conclusion du test ?

6. Rappeler la définition de la puissance statistique. À l'aide de celle-ci, que peut-on dire de ce test ?
7. Montrer qu'en tant qu'estimateur de N , $\hat{N}_n^{(\max)}$ est biaisé.
8. Donner un estimateur non biaisé de N .

Exercice 3. *Taux de réussite (extrait de l'examen de mai 2017)*

Dans une unité d'enseignement de Licence 2, on veut savoir si les N étudiants de l'année en cours ont réussi dans des proportions similaires à celles attendues par le ministère. On modélise le succès ou l'échec du i -ème étudiant par une variable aléatoire X_i de loi de Bernoulli de paramètre $p \in]0, 1[$. On suppose que les variables aléatoires $X_1, \dots, X_N$ sont indépendantes. On pose

$$T = \sum_{i=1}^N X_i$$

le nombre d'étudiants ayant réussi cette année.

1. Quelle est la loi de T ?

On suppose que l'on a $N = 100$ étudiants et que le ministère souhaiterait un taux de succès $p_0 = 0.9$. On donne ci-dessous le tableau de la fonction de répartition d'une variable T_0 de loi binomiale $\mathcal{B}(N, p_0)$ de paramètres N et p_0 .

k	81	82	83	84	85	86	87	88	89	90
$100 \times P(T_0 \leq k)$	0.46	1.00	2.06	3.99	7.26	12.39	19.82	29.70	41.68	54.49
k	91	92	93	94	95	96	97	98	99	100
$100 \times P(T_0 \leq k)$	67.91	79.39	88.28	94.42	97.63	99.22	99.91	99.97	99.997	100

2. Construire un test de niveau $\alpha = 0.05$ de l'hypothèse nulle $H_0 : \ll p = p_0 \gg$ (les étudiants ont réussi dans des proportions en accord avec celles demandées par le ministère) contre l'hypothèse $H_1 : \ll p \neq p_0 \gg$, à partir de la statistique de test T .
3. Cette année, on a observé que 84 étudiants ont réussi l'unité d'enseignement. Que peut-on en conclure ?
4. L'année d'avant, 95 étudiants avaient réussi. Peut-on dire que cela était en accord avec les recommandations du ministère ($p = p_0$) au niveau $\alpha = 0.05$?

Exercice 4. *Le dé est-il truqué ? (extrait de l'examen de juin 2019)*

Sarah et Jade veulent jouer avec un dé à six faces. Avant de commencer une partie, Jade, méfiante, cherche à déterminer si le dé choisi par Sarah est bien équilibré, ou bien s'il s'agit de son dé « truqué » qui tombe un peu trop souvent sur de petits chiffres.

1. Sarah assure à Jade que le dé est bien équilibré. Pour le vérifier, Jade lance 50 fois le dé et note uniquement le nombre de fois N où le dé tombe sur la face 1 ou sur la face 2. Construire un test de l'hypothèse nulle « le dé est équilibré » contre l'hypothèse alternative « le dé est truqué », au niveau 1%, fondé sur le résultat de l'expérience menée par Jade. On pourra utiliser le tableau suivant qui donne certaines valeurs de la fonction de répartition de N lorsque le dé est équilibré (P_{equ}) et lorsqu'il s'agit du dé truqué (P_{tru}).

k	15	16	17	18	19	20	21	22	23	24
$P_{equ}(N \leq k)$	0.369	0.487	0.605	0.713	0.804	0.874	0.924	0.958	0.978	0.989
$P_{tru}(N \leq k)$	0.003	0.008	0.016	0.032	0.059	0.101	0.161	0.24	0.336	0.444

k	25	26	27	28	29	30	31	32	33	34
$P_{equ}(N \leq k)$	0.995	0.998	0.999	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1
$P_{tru}(N \leq k)$	0.556	0.664	0.760	0.839	0.899	0.941	0.968	0.984	0.992	0.997

2. Calculer la puissance du test construit à la question 1. Commenter la valeur obtenue.
3. Jade, après 50 lancers, obtient 10 fois la face 1 et 16 fois la face 2. Quelle conclusion peut-elle en tirer ? Doit-elle accuser son amie Sarah de lui avoir menti ?

Exercice 5. *Surréservation dans un hôtel*

Un hôtel sur l'île de Ré propose à la réservation 200 chambres lors du week-end de la Pentecôte. Le propriétaire sait qu'environ 20% des réservations ne sont pas confirmées ou sont annulées (hors épidémie de coronavirus). Souhaitant s'assurer que son hôtel sera plein durant cette période cruciale, il va laisser un nombre $n \geq 200$ réservations s'effectuer.

On notera X la variable aléatoire comptant, parmi les n possibles, les personnes ayant confirmé leur réservation pour ce week-end.

1. Quelle est la loi suivie par X ?
2. Le propriétaire souhaite éviter les mauvais commentaires auxquels il peut s'exposer en cas de surréservation.

Quel est le nombre maximum de réservations que le propriétaire peut accepter pour être sûr à 95% qu'il n'y aura pas de clients qui, au moment d'arriver, devront être relogés pour cause d'hôtel déjà plein ?

Indication : On cherche à s'assurer que $P(X > 200) \leq 0.05$.

3. Même question dans le cas d'un hôtel sur une île tropicale, avec 2000 chambres, mais avec un risque de voir une réservation non confirmée de 10% et un gérant qui souhaite être sûr à 99% d'éviter d'avoir à refuser des clients ayant réservé une chambre, car il doit alors les indemniser.

Exercice 6. *Étude de marché*

Un vendeur de pommes bio souhaite s'implanter sur l'un des sites de l'Université Paris-Saclay, où il y a environ $N = 10000$ étudiants. Avant cela, il fait un sondage auprès de $n = 250$ étudiants pour se rendre compte des chances pour son commerce d'être rentable. Il faut pour cela qu'il vende 300 pommes par jour. Parmi les n étudiants interrogés, $n_1 = 50$ lui assurent que le concept leur plaît, et qu'ils lui achèteront une pomme par jour.

1. Modéliser la situation du commerçant afin de déterminer un intervalle de confiance, à 90%, pour le nombre de pommes vendues quotidiennement.
2. Le sondage est-il utile au commerçant pour conclure quant à la rentabilité de ses pommes bio ?
3. Refaire l'estimation pour $n_1 = 10$: quels sont les changements ?

Exercice 7. *Promenade en forêt (extrait de l'examen de mai 2017)*

Vincent, qui habite dans les Vosges, aime se promener en forêt et observer les oiseaux. Il s'intéresse en particulier à deux espèces de pics, qu'il observe souvent : le pic épeiche et le pic noir (on supposera qu'il n'observe pas d'autre espèce de pic). Il aimerait déterminer la proportion p de pics noirs parmi l'ensemble des pics.

1. Énoncer de façon **précise** l'inégalité de Bienaymé-Tchebychev.
2. Soit $X_1, X_2, \dots, X_n$ des variables aléatoires indépendantes et de même loi, d'espérance $M \in \mathbb{R}$ et de variance $V \in [0, +\infty[$. Soit $\varepsilon > 0$ un réel.

(a) Montrer que

$$P(|X_1 + \dots + X_n - nM| \geq n\varepsilon) \leq \frac{V}{n\varepsilon^2}.$$

(b) Déterminer un intervalle $]a(n, \varepsilon); b(n, \varepsilon)[$ tel que

$$P(X_1 + \dots + X_n \in]a(n, \varepsilon); b(n, \varepsilon)[) \geq 1 - \frac{V}{n\varepsilon^2}.$$

(c) En déduire un intervalle $\hat{J}$, fonction de $X_1, \dots, X_n$, n et ε uniquement, tel que

$$P(M \in \hat{J}) \geq 1 - \frac{V}{n\varepsilon^2}.$$

Bien préciser les bornes de $\hat{J}$.

3. Vincent va se promener dans les forêts de sa région et note, pour chaque pic qu'il observe, s'il est noir ou épeiche. Vincent a l'impression qu'il y a deux fois plus de pics épeiches que de pics noirs. Dans cette question 3 uniquement, on suppose que ceci correspond bien à la réalité.
 - (a) On note A la variable aléatoire qui vaut 1 si le premier pic observé par Vincent est noir, et qui vaut 0 sinon. Quelle est la loi de A ? Donner les valeurs de son espérance et de sa variance.
 - (b) En utilisant la question 2, donner une minoration de la probabilité que Vincent observe entre 9 et 15 pics noirs (au sens large) parmi les 36 premiers pics observés.
4. Vincent pense désormais que son estimation initiale n'est pas exacte. Il retourne en forêt et note le nombre S_n de pics noirs parmi les n premiers pics qu'il observe. Il propose d'estimer p par $\hat{p}_n = S_n/n$.

- (a) Calculer l'espérance de $\hat{p}_n$. En déduire une propriété de $\hat{p}_n$ comme estimateur de p .
 - (b) Déduire de la question 2 un intervalle d'amplitude 10% qui contient « souvent » $\hat{p}_n$, ainsi qu'une minoration de la probabilité que cet intervalle contienne $\hat{p}_n$.
5. Vincent est maintenant certain que moins d'un quart des pics sont des pics noirs. Il souhaite affiner encore son estimation.
- (a) Soit B la variable qui vaut 1 si le premier pic observé est noir (lors de la prochaine sortie de Vincent), et qui vaut 0 sinon. À l'aide de l'hypothèse de Vincent, majorer la variance de B .
 - (b) À l'aide de $\hat{p}_n$, construire un intervalle de confiance pour p de probabilité de couverture 95% (au moins).
 - (c) Lors de ses promenades en forêt, Vincent note le nombre de pics noirs pour chaque série de 20 pics observés :

4, 7, 5, 3, 4, 3, 3, 6, 2, 5, 6, 3, 1, 2, 4, 6, 5, 6, 6, 5.

Au vu de ces résultats, proposer un intervalle contenant la proportion de pics noirs, avec un risque 5%. Est-ce cohérent avec l'hypothèse de Vincent ?

- (d) Combien de pics Vincent doit-il observer pour que l'intervalle de confiance construit à la question 5b soit d'amplitude au plus 5% ?

Exercice 8. *Pannes d'ascenseur (extrait de l'examen de mai 2019)*

Suzanne est réparatrice d'ascenseurs. Sa longue expérience l'a amenée à modéliser la loi du nombre de pannes qu'elle doit réparer chaque jour par une loi de Poisson.

1. Sur la période 2009–2018, Suzanne estime qu'il y avait en moyenne 2 pannes par jour (parmi ses clients). En supposant cette estimation correcte, que vaut le paramètre λ_0 de la loi de Poisson qui décrit le nombre quotidien de pannes sur ces dix dernières années ?
2. Soit Z le nombre de pannes que Suzanne doit réparer sur une période de 30 jours. En supposant l'estimation précédente correcte, quelle est la loi de Z ? On précisera les éventuelles hypothèses de modélisation faites pour répondre à la question.
3. Énoncer de façon précise l'inégalité de Bienaymé-Tchebychev.
4. On rappelle que, pour tout $\mu > 0$, la variance de la loi de Poisson de paramètre μ est égale à μ . À l'aide de la question précédente, construire un intervalle (déterministe), de la forme $[a; b]$, qui contient Z avec probabilité au moins 95%.
5. Au premier janvier 2019, la marque d'ascenseurs dont Suzanne est spécialiste a mis à jour le logiciel gérant tous ses ascenseurs. Elle se demande si cette mise à jour n'a pas modifié la fréquence des pannes. Construire un test de niveau 5% permettant de répondre à cette question, en utilisant comme statistique de test la variable aléatoire S qui compte le nombre de pannes observées du 1er au 30 janvier 2019 (inclus).
6. Il y a eu 30 pannes entre le 1er et le 30 janvier 2019. Que peut-on conclure ?

Exercice 9. *Dépistage de problèmes de vision nocturne (extrait de l'examen de mai 2012)*

1. Énoncer de façon **précise** l'inégalité de Bienaymé-Tchebychev.
2. Soit $X_1, X_2, \dots, X_n$ n variables aléatoires indépendantes, d'espérance m et de variance σ^2 . On note S_n la somme $X_1 + \dots + X_n$.

- (a) Calculer $E[S_n]$.
 - (b) Montrer que la variance de S_n vaut $n\sigma^2$.
 - (c) Soit $\varepsilon > 0$ un réel, **montrer** que : $P(|S_n - nm| \geq n\varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$.
3. Le service de médecine préventive d'une région veut faire une étude sur les problèmes de vision nocturne. Un test a été mis au point. D'après les retours d'une autre région utilisant ce test, il est clair que moins de 15% des personnes interrogées échoueront au test. Le responsable du service voudrait faire une estimation de la proportion p des habitants de cette région qui échoue à ce test en donnant un intervalle de longueur 6%. Il voudrait être sûr du résultat obtenu avec une probabilité de se tromper inférieure à 5%. Il se demande à combien d'habitants il faut faire passer le test. Étant donné qu'il y a plusieurs centres de médecine préventive dans cette région et que les personnes convoquées sont d'origines suffisamment diverses, on peut considérer que les tests sont indépendants et identiquement distribués. On notera n le nombre de personnes à convoquer.
- (a) On fait faire le test à une personne. On note X la variable aléatoire qui vaut 1 si elle échoue au test et 0 sinon. Quelle loi suit X ?
 - (b) Quelle hypothèse peut-on faire sur p et donc sur la variance de X ?
 - (c) On note n le nombre de personnes à convoquer et S_n le nombre de personnes qui ont échoué au test. En utilisant la question 2 (expliquer pourquoi on peut l'utiliser ici), majorer $P(|\frac{S_n}{n} - p| \geq \frac{3}{100})$ par une constante m_n qui ne dépend que de n .
 - (d) Déterminer un n (le plus petit possible) tel que $m_n \leq 5\%$.
 - (e) On fait exactement 3000 tests, on trouve 162 personnes qui ont échoué au test. Quelles inégalités déduit-on pour p de la question (d) ?
 - (f) Il est actuellement admis que si 7% de la population a des problèmes de vision nocturne, il faut faire une campagne de prévention. Quelle conclusion le responsable peut-il tirer du résultat précédent ?

Exercice 10. Essai thérapeutique

La probabilité de guérison d'une certaine maladie avec le traitement habituel est de $\frac{1}{2}$. Un laboratoire annonce la sortie d'un médicament qui, ajouté au traitement habituel, guérirait cette maladie avec une probabilité $\frac{3}{4}$. Un hôpital procède à des essais thérapeutiques sur 12 patients traités avec ce nouveau médicament en plus du traitement habituel, afin de tester les deux hypothèses suivantes : H_0 : « le nouveau médicament n'a aucun effet » et H_1 : « le nouveau médicament a l'effet annoncé ».

1. Construire un test de niveau 5% de l'hypothèse H_0 contre l'hypothèse H_1 .
2. Calculer l'erreur de deuxième espèce et la puissance du test. Donner leurs significations pour l'hôpital.
3. Sur les 12 patients traités avec ce nouveau médicament en plus du traitement habituel, 9 guérissent.
On veut que la probabilité de se tromper, lorsque le médicament n'a aucun effet, soit inférieure à 5%. Que peut-on conclure ?
4. Reprendre les questions précédentes dans le cas où l'on a traité 36 patients avec le nouveau médicament. Que constate-t-on sur la puissance du test ? Cette constatation vous étonne-t-elle ? Quelle sera la conclusion du test si, sur les 36 patients traités, 27 ont guéri ?

Valeurs approchées de $F_X(k) = P(X \leq k)$ pour $X \sim B(n, p)$

$n = 12, p = 0.5$

k	4	5	6	7	8	9	10	11	12
$F_X(k)$	0.193	0.387	0.613	0.806	0.927	0.981	0.997	0.999	1

$n = 12, p = 0.75$

k	4	5	6	7	8	9	10	11	12
$F_X(k)$	0.278e-2	0.143e-1	0.544e-1	0.158	0.351	0.609	0.842	0.968	1

$n = 36, p = 0.5$

k	22	23	24	25	26	27	28	29	30
$F_X(k)$	0.934	0.967	0.986	0.994	0.998	0.999	1	1	1

$n = 36, p = 0.75$

k	22	23	24	25	26	27	28	29	30
$F_X(k)$	0.461e-1	0.922e-1	0.167	0.275	0.412	0.564	.710	0.832	0.917

Exercice 11.

Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes de loi de Bernoulli de paramètre p , et $Y_1, \dots, Y_n$ des variables aléatoires indépendantes de loi de Bernoulli de paramètre q et indépendantes de $X_1, \dots, X_n$.

1. Quelle est la loi de $X_1 Y_1$?
2. Montrer que $V_n = \frac{1}{n} \sum_{i=1}^n X_i Y_i$ est un estimateur sans biais de pq . Quelle est sa variance ?
3. Proposer (en justifiant) un estimateur empirique U_n de p et un estimateur empirique T_n de q .
4. Montrer que $R_n = U_n T_n$ est un estimateur sans biais de pq .
5. Montrer que $|R_n - pq| \leq |U_n - p| + |T_n - q|$.
6. Montrer que pour tout $\epsilon > 0$,

$$P(|R_n - pq| \geq \epsilon) \leq P\left(|U_n - p| \geq \frac{\epsilon}{2}\right) + P\left(|T_n - q| \geq \frac{\epsilon}{2}\right).$$

En déduire que pour tout $\epsilon > 0$,

$$\lim_{n \rightarrow +\infty} P(|U_n T_n - pq| \geq \epsilon) = 0.$$

7. Montrer que

$$\text{Var}(R_n) = (\text{Var}(U_n) + p^2)(\text{Var}(T_n) + q^2) - p^2 q^2,$$

et en déduire que

$$\text{Var}(V_n) - \text{Var}(R_n) = \frac{pq(1-p)(1-q)(n-1)}{n^2}.$$

Qu'en pensez-vous ?