VIII- Tests d'hypothèses sur les valeurs des paramètres d'une variable aléatoire

Dans ce chapitre, $X_1, \ldots, X_n$ désigne un échantillon aléatoire d'une loi qui dépend d'un paramètre réel inconnu θ.

Considérons l'hypothèse sur le paramètre θ :

$$(H) : \theta = \theta_0$$

où θ_0 est un valeur explicite.

On veut construire un test qui utilise un échantillon expérimental $(x_1, \ldots, x_n)$ pour éprouver cette hypothèse. On procède de la façon suivante.

On choisit un estimateur Θ du paramètre θ. On note θ_e sa valeur expérimentale.

On se donne une *variable aléatoire discriminante* de la forme $D = \delta(X_1, \ldots, X_n, \theta_0)$, où δ est une fonction déterministe à valeurs réelles. On fait en sorte que sa valeur expérimentale $d_e = \delta(x_1, \ldots, x_n, \theta_0)$ permette de comparer les valeurs θ_0 et θ_e, en en reflétant par exemple la distance. On se donne aussi une *zone de rejet* R ($R \subset \mathbb{R}$), et on décide :

- de rejeter l'hypothèse (H) si $d_e \in R$
- de considérer que l'expérience ne contredit pas (H) sinon

En prenant cette décision, on court un risque de se tromper.

Si on est dans la deuxième situation, la formulation de la conclusion est tellement molle qu'on ne court pas grand risque de se tromper. (Ce qui ne veut pas dire pour autant que considérer qu'une expérience ne contredit pas (H) est toujours anodin. Supposons par exemple que (H) signifie "Le fonctionnement de la centrale nucléaire de Blaye est normal"...)

Si on est dans la première des situations, il se peut que le paramètre θ soit vraiment égal à θ_0, et que le fait que d_e soit dans la zone de rejet R soit un fait de hasard. Si c'est le cas, le test nous fait rejeter (H) à tort, et plus précisément, si θ_0 est la vraie valeur de θ, en utilisant ce test, on se trompera dans la décision (environ) $100\,\alpha$ fois sur 100, où :

$$\alpha = P\{ D \in R \}.$$

Ce nombre α , ou le pourcentage $(100\,\alpha)\%$, s'appelle le *niveau de risque du test* . Pour pouvoir calculer ce risque, il faut donc choisir la fonction discriminante D de telle sorte que sa loi soit connue lorsque (H) est vraie.

Par la suite, pour construire un test de (H), on fixera en général dès le départ le niveau de risque α, et on définira la zone de rejet R_α de sorte que $P\{ D \in R_\alpha \} = \alpha$.

1- Valeur de l'espérance d'une variable normale de variance connue

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, où σ est connu et μ inconnu. On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons μ_0 , et soit à tester l'hypothèse :

$$(H) \ : \ \mu = \mu_0$$

Construisons le test.

On utilise l'estimateur $\overline{X}$ du paramètre μ. On note $\overline{x}$ sa valeur expérimentale.
Fixons à α $(0 < \alpha < 1)$ le niveau de risque du test. (Comme α mesure un risque de se tromper, on choisit α "petit", par exemple $\alpha = 0{,}05$ ou $0{,}1$).

Si l'hypothèse (H) est vraie, la variable aléatoire $D = \dfrac{\sqrt{n}\ (\overline{X} - \mu_0)}{\sigma}$ suit la loi $\mathcal{N}(0 \ ; \ 1)$.

Nous choisirons cette variable aléatoire comme variable discriminante. Remarquons que sa valeur expérimentale $d_e = \dfrac{\sqrt{n}\ (\overline{x} - \mu_0)}{\sigma}$ reflète bien la distance entre $\overline{x}$, la valeur expérimentale du paramètre μ, et μ_0, la valeur à tester.

Définissons $t_{\alpha/2}$ par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \ , \quad Y \text{ suivant la loi } \mathcal{N}(0 \ , \ 1).$$

et la zone de rejet :

$$R_\alpha = \{ \ d \in \mathbb{R} \ / \ |d| > t_{\alpha/2} \ \}$$

La construction du test est achevée.

La mise en œuvre de ce test au niveau de risque α consiste à décider de :

- rejeter l'hypothèse (H) si $\left| \dfrac{\sqrt{n}\ (\overline{x} - \mu_0)}{\sigma} \right| > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon .

<u>Exercice 8-1</u> : On suppose que lorsqu'un signal de valeur μ est émis d'un point A, la valeur du signal reçu au point B est bruitée et suit une loi normale $\mathcal{N}(\mu \ , \ 2)$. Une personne au point B s'attend à ce que le signal émis ait la valeur 8. Or, le même signal est émis 5 fois du point A, et la valeur moyenne reçue au point B est 9,5. Cette personne doit-elle remettre en cause son hypothèse ?

L'hypothèse $(\mu = \mu_0)$ dont nous venons de décrire le test est ce qu'on appelle une *hypothèse simple*, car, sous cette hypothèse, la loi de l'échantillon est complètement déterminée.
Soit maintenant à tester l'hypothèse *composite* :

$$(H) : \mu \leq \mu_0$$

où μ_0 est une valeur explicite du paramètre.
Construisons-en un test de niveau α ($0 < \alpha < 1$).

On utilise l'estimateur $\bar{X}$ du paramètre μ.

On décidera de rejeter (H) lorsque la valeur de $\bar{x}$ est trop grande par rapport à μ_0, ou, ce qui revient au même, lorsque $d_e = \dfrac{\sqrt{n}\ (\bar{x} - \mu_0)}{\sigma} > c$ pour un certain c.

Pour construire c en fonction du niveau α, supposons que l'hypothèse (H) est vraie, et plus précisément, supposons que μ ($\mu \leq \mu_0$) est la vraie valeur du paramètre. Le risque de rejeter à tort (H) est alors quantifié par $P\{ \dfrac{\sqrt{n}\ (\bar{X} - \mu_0)}{\sigma} > c\}$. Or, $\dfrac{\sqrt{n}\ (\bar{X} - \mu)}{\sigma}$ suit la loi $\mathcal{N}(0, 1)$ et ce risque est donc :

$$P\{ \frac{\sqrt{n}\ (\bar{X} - \mu_0)}{\sigma} > c\} = P\{ Y > c + \frac{\sqrt{n}\ (\mu_0 - \mu)}{\sigma} \}$$

où Y suit la loi $\mathcal{N}(0 , 1)$. Il est le plus grand lorsque $\mu = \mu_0$, et il vaut alors $P\{ Y > c \}$. On va donc choisir c tel que cette probabilité soit égale à α. Ainsi, on saura que si l'hypothèse (H) est vérifiée, le test rejettera a tort (H) au plus (environ) $100\ \alpha$ fois sur 100.

En résumé, la mise en œuvre de ce test au niveau de risque α consiste à décider de :

- rejeter l'hypothèse (H) si $\dfrac{\sqrt{n}\ (\bar{x} - \mu_0)}{\sigma} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où t_α est défini par :

$$P(Y > t_\alpha) = \alpha \quad , \quad Y \text{ suivant la loi } \mathcal{N}(0 , 1).$$

2- Valeur de l'espérance d'une variable normale de variance inconnue

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, où μ et σ sont inconnus. On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons μ_0 , et soit à tester l'hypothèse :

$$(H) : \mu = \mu_0$$

Construisons-en un test au niveau de risque α ($0 < \alpha < 1$).

On utilise les estimateurs $\overline{X}$ et S de paramètre μ et σ. On note $\overline{x}$ et s leurs valeurs expérimentales.

Si l'hypothèse (H) est vraie, la variable aléatoire $D = \frac{\sqrt{n}\ (\overline{X} - \mu_0)}{S}$ suit la loi de Student à n-1 degrés de liberté. Remarquons qu'il est moins clair que dans le cas où σ est connu que sa valeur expérimentale $\frac{\sqrt{n}\ (\overline{x} - \mu_0)}{s}$ reflète la distance entre $\overline{x}$ et μ_0, puisque le dénominateur s dépend de la valeur expérimentale $(x_1, \ldots, x_n)$. Il est pourtant d'usage de choisir cette variable aléatoire D comme variable discriminante.
Nous définirons alors $t_{\alpha/2}$ par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \ , \quad Y \text{ suivant la loi de Student } t_{n-1} .$$

et la zone de rejet :

$$R_\alpha = \{ d \in \mathbb{R} \ / \ |d| > t_{\alpha/2} \ \}$$

La mise en œuvre du test au niveau de risque α consiste donc à décider de :

- rejeter l'hypothèse (H) si $\left| \frac{\sqrt{n}\ (\overline{x} - \mu_0)}{s} \right| > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon .

On pourrait, de même que dans le cas de la variance connue, construire un test au niveau de risque a de l'hypothèse composite ($\mu \leq \mu_0$).

Exercice 8-2 : L'utilisateur d'un certain câble exige que sa charge moyenne de rupture soit au moins de 200 tonnes. Il a testé 8 de ces câbles et trouvé les charges de rupture :

210 195 197,4 199 198 202 196 195,5

On suppose que la charge de rupture d'un câble suit une loi normale.
Que conclure, au niveau de risque de 5% ? Au niveau de risque de 10% ?

3- Valeur de la variance d'une variable normale

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi $\mathcal{N}(\mu, \sigma)$, où μ et σ sont inconnus. On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons μ_0 , et soit à tester l'hypothèse :

$$(H) \ : \ \sigma = \sigma_0$$

Construisons-en un test au niveau de risque α $(0 < \alpha < 1)$.

On utilise les estimateurs $\bar{X}$ et S de paramètre μ et σ. On note $\bar{x}$ et s leurs valeurs expérimentales.

Si l'hypothèse (H) est vraie, la variable aléatoire, $\frac{(n-1)S^2}{\sigma_0^2}$ suit la loi du khi-deux à n-1 degrés de liberté. Nous la choisissons comme variable discriminante. Sa valeur expérimentale $\frac{(n-1)s^2}{\sigma_0^2}$ est fonction de s et permet donc la comparaison de s et σ_0.

On définit la zone de rejet :

$$R_\alpha = \{ d \in \mathbb{R} \ / \ d > t_{\alpha/2} \ \text{ ou } \ d < t_{1-\alpha/2} \ \}$$

avec :

$$P(Y > t_\beta) = \beta \ , \quad Y \text{ suivant la loi } \chi^2_{n-1}$$

La construction du test est achevée.

La mise en œuvre de ce test au niveau de risque α consiste à décider de :

- considérer que l'expérience ne contredit pas (H) si $t_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma_0^2} < t_{\alpha/2}$,
- rejeter l'hypothèse (H) sinon.

En suivant une démarche analogue à celle décrite dans le paragraphe 1, on peut justifier l'utilisation, pour tester l'hypothèse composite ($\sigma \leq \sigma_0$) au risque α, du test qui consiste à

- rejeter l'hypothèse (H) si $\frac{(n-1)s^2}{\sigma_0^2} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon.

Si l'espérance μ est connue, on construit les tests de manière analogue, en utilisant l'estimateur de la variance $S'^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$ et en remarquant que si (H) est vraie, $D = \frac{nS'^2}{\sigma_0^2}$ suit la loi du khi-deux à n degrés de liberté.

Exercice 8-3 : Le système de mesure d'une pompe à essence est tel que le nombre de litres affichés suit une loi normale d'espérance égale au nombre de litres distribués et d'écart-type inconnu σ. Ce système est considéré comme efficace si σ est inférieur à 0,075 litres. Par 20 mesures indépendantes, on a testé un système nouvellement installé et obtenu l'estimation $s^2 = 0{,}00625$. Le système de mesure est-il efficace ?

4- Valeur de la probabilité d'un évènement

Supposons que $X_1, \ldots, X_n$ est un échantillon aléatoire de loi de Bernoulli $\mathfrak{B}(p)$, où p est inconnu.
On suppose qu'on dispose d'une valeur expérimentale $(x_1, \ldots, x_n)$ de cet échantillon.

Fixons p_0 , et soit à tester l'hypothèse :

$$(H) \ : \ p = p_0$$

Construisons-en un test au niveau de risque α $(0 < \alpha < 1)$.

On utilise comme estimateur de p la moyenne de l'échantillon, $\bar{X}$. On note p_e la valeur expérimentale correspondante.

Si l'hypothèse (H) est vraie, la variable aléatoire $n\bar{X}$ suit la loi du binôme $\mathfrak{B}(n, p_0)$. Nous la choisissons comme variable discriminante. Par un calcul itératif, - ou en utilisant des tables ou abaques -, on peut déterminer $k^-_{\alpha/2}$ et $k^+_{\alpha/2}$ tels que :

$$k^-_{\alpha/2} = \max \{ k \ / \ \sum_{i=0}^{k} C_n^i \ {p_0}^i (1 - p_0)^{n-i} \leq \alpha/2 \ \}$$

$$k^+_{\alpha/2} = \min \{ k \ / \ \sum_{i=k}^{n} C_n^i \ {p_0}^i (1 - p_0)^{n-i} \leq \alpha/2 \ \}$$

La mise en œuvre de ce test au niveau de risque α consiste alors à décider de :

- considérer que l'expérience ne contredit pas (H) si $k^-_{\alpha/2} < np_e < k^+_{\alpha/2}$,
- rejeter l'hypothèse (H) sinon.

Supposons maintenant que la taille n de l'échantillon est assez grande pour qu'on puisse, sous l'hypothèse (H), donner une bonne approximation de la loi de $\dfrac{\sqrt{n}\ (\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}}$ par la loi $\mathfrak{N}(0 ; 1)$. (Il est d'usage de considérer que cette approximation est très bonne lorsque $np_0(1 - p_0) \geq 10$). On peut alors proposer un test au niveau de risque α de mise en œuvre beaucoup plus simple.

On choisit $\dfrac{\sqrt{n}\ (\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}}$ comme variable discriminante. Définissant $t_{\alpha/2}$ par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \ , \quad Y \text{ suivant la loi } \mathfrak{N}(0 , 1),$$

on a :

$$P\{ \left| \frac{\sqrt{n}\ (\bar{X} - p_0)}{\sqrt{p_0(1-p_0)}} \right| > t_{\alpha/2} \} \approx \alpha \ .$$

La mise en œuvre de ce test au niveau de risque α consiste donc à décider de :

- rejeter l'hypothèse (H) si $\left| \dfrac{\sqrt{n}\ (p_e - p_0)}{\sqrt{p_0(1-p_0)}} \right| > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon .

Exercice 8-4 : La chaîne de fabrication de montres est conçue pour qu'au plus 2% des montres soient défectueuses. Sur 500 montres testées, on en a trouvé 16 défectueuses. Doit-on conclure à un dysfonctionnement de la chaîne de fabrication ? (Proposer et utiliser un test unilatéral).

5- Valeur de l'espérance d'une variable aléatoire de loi quelconque

Supposons que $X_1, \ldots, X_n$ est un échantillon aléatoire de loi quelconque et qu'on veuille tester une hypothèse sur l'espérance μ de sa loi.

Si le type de la loi de l'échantillon est connue, il faut en principe faire une analyse analogue à celle que nous avons faite pour la loi normale : choisir un estimateur ($\bar{X}$ n'est pas forcément le meilleur : voir le chapitre VII §4 ...), choisir une fonction discriminante (de loi connue ou calculable, c'est là le plus gros problème...), etc…

Cependant, pour des valeurs de n assez grandes, et si l'écart-type σ des X_i est connu, on sait, d'après le théorème central-limite, que sous l'hypothèse (H), $\frac{\sqrt{n}\ (\bar{X} - \mu_0)}{\sigma}$ suit approximativement la loi $\mathcal{N}(0 ; 1)$. On pourra alors construire un test comme on l'a fait dans le paragraphe 4 sur la loi de Bernoulli. Si l'écart-type σ des X_i est inconnu, on utilise généralement la variable discriminante $\frac{\sqrt{n}\ (\bar{X} - \mu_0)}{S}$, en considérant qu'elle suit approximativement la loi $\mathcal{N}(0 ; 1)$, mais on ne peut pas le justifier dans un cadre général. Dans tous les cas, il faut remarquer que si le type de loi des X_i est inconnu, on ne sait pas pour quelles valeurs de n ces approximations sont valides. On ne se risquera pas à utiliser de tels tests si n est plus petit que 30.

6- Intervalle de confiance pour l'estimation d'un paramètre

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une loi qui dépend d'un paramètre réel inconnu θ, et soit Θ un estimateur du paramètre θ. On suppose disposer d'un échantillon expérimental $(x_1, \ldots, x_n)$, et on note θ_e sa valeur expérimentale. On suppose enfin qu'on dispose d'un test de niveau de risque donné α $(0 < \alpha < 1)$ pour tester les hypothèses $(\theta = \theta_0)$.
On définit alors l'intervalle de confiance au niveau de confiance $(1- \alpha)$ de l'estimation du paramètre θ comme l'ensemble $I_{1-\alpha}$ des valeurs θ_0 qui ne sont pas rejetées par ce test.

En utilisant les tests proposés dans les paragraphes précédents, on obtient les intervalles de confiance au niveau (1- α) suivants :

- Intervalle de confiance de l'espérance d'une variable normale d'écart-type connu σ :

$$I_{1-\alpha} = \left[\bar{x} - t_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + t_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

avec $t_{\alpha/2}$ défini par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \qquad \text{où Y suit la loi } \mathcal{N}(0 , 1).$$

- Intervalle de confiance de l'espérance d'une variable normale de variance inconnue :

$$I_{1-\alpha} = \left[\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \quad , \quad \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

avec $t_{\alpha/2}$ défini par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \qquad \text{où Y suit la loi de Student } t_{n-1}.$$

- Intervalle de confiance de la variance d'une variable normale d'espérance inconnue :

$$I_{1-\alpha} = \left[\frac{(n-1)s^2}{t_{\alpha/2}} \quad , \quad \frac{(n-1)s^2}{t_{1-\alpha/2}} \right]$$

avec :

$$P(Y > t_{\beta}) = \beta \qquad \text{où Y suit la loi } \chi^2_{n-1} .$$

- Intervalle de confiance de la variance d'une variable normale d'espérance connue :

$$I_{1-\alpha} = \left[\frac{ns'^2}{t_{\alpha/2}} \quad , \quad \frac{ns'^2}{t_{1-\alpha/2}} \right]$$

avec :

$$P(Y > t_{\beta}) = \beta \qquad \text{où Y suit la loi } \chi^2_{n} .$$

- Intervalle de confiance du paramètre d'une variable de Bernoulli pour les grandes valeurs de n :

$$I_{1-\alpha} \approx \left[p_e - t_{\alpha/2} \sqrt{\frac{p_e(1-p_e)}{n}} \quad , \; p_e + t_{\alpha/2} \sqrt{\frac{p_e(1-p_e)}{n}} \right]$$

avec $t_{\alpha/2}$ défini par :

$$P(Y < - t_{\alpha/2}) = P(Y > t_{\alpha/2}) = \alpha /2 \qquad \text{où Y suit la loi } \mathcal{N}(0 , 1).$$

Exercice 8-5 : On suppose que lorsqu'un signal de valeur μ est émis d'un point A, la valeur du signal reçu au point B est bruitée et suit une loi normale $\mathcal{N}(\mu , 2)$.

a) Pour réduire l'erreur de transmission, on envoie le même signal 9 fois. Les valeurs reçues sont :

5 8,5 12 15 7 9 7,5 6,5 10,5 .

Quel est l'intervalle de confiance bilatéral de la valeur émise μ, au niveau de confiance 0,95 ?

b) Combien de fois le même signal doit-il être envoyé pour que l'intervalle de confiance de μ au niveau 0,95 soit de demi-longueur inférieure à 0,1 ?

Si on dispose d'un test unilatéral de niveau de risque donné α ($0 < \alpha < 1$) pour tester les hypothèses ($\theta \leq \theta_0$), on définit l'intervalle de confiance $]\ldots, +\infty[$ au niveau de confiance $(1- \alpha)$ de l'estimation du paramètre θ comme l'ensemble $I_{1-\alpha}$ des valeurs θ_0 qui ne sont pas rejetées par ce test.

7- Exercices

Exercice 8-6 : Un procédé de fabrication exige d'une certaine solution chimique d'avoir un pH exactement égal à 8,20. La méthode de mesure de pH utilisée donne un résultat qui suit la loi normale d'écart-type 0,02 et d'espérance égale à la vraie valeur du pH. On a mesuré 10 fois le pH de la solution et trouvé :

8,18 8,16 8,17 8,22 8,19 8,17 8,15 8,21 8,16 8,18

a) Que conclure au niveau de risque de 5% ?

b) Que conclure au niveau de risque de 5‰ ?

Exercice 8-7 : On a constaté que sur n = 100 naissances, g = 49 ont été des naissances de garçons. Est-il raisonnable d'admettre que les naissances sont égalements réparties entre garçons et filles ? Même question pour 490 naissances de garçons sur un total de 1 000, de 4 900 sur un total de 10 000.

Exercice 8-8 : Reprendre les données et les questions de l'exercice 8-5, mais en supposant que lorsqu'un signal de valeur μ est émis d'un point A, la valeur du signal reçu au point B suit une loi normale $\mathcal{N}(\mu , \sigma)$, avec μ et σ inconnus.

Exercice 8-9 : Un procédé de vérification de l'épaisseur de rondelles métalliques fournit une mesure qui suit une loi normale d'espérance égale à la vraie valeur de l'épaisseur et d'écart-type inconnu. On a mesuré 10 fois l'épaisseur d'une rondelle et trouvé :

1,23 1,24 1,26 1,20 1,30 1,33 1,25 1,28 1,24 1,26 mm.

Quel est l'intervalle de confiance au niveau 0,8 de l'écart-type de l'épaisseur d'une rondelle ?

Exercice 8-10 : Dans une population africaine isolée, on a testé 72 personnes choisies au hasard, et observé que 9 d'entre elles portent une anomalie génétique particulière. Quelle est l'intervalle de confiance au niveau 0,95 de la fréquence de cette anomalie dans la population ?

Exercice 8-11 : Entre le premier et second tour des élections présidentielles, un candidat C commande à un institut de sondage une évaluation de ses chances de gagner. Sur 1000 personnes interrogées et ayant l'intention d'exprimer leur suffrage, 515 déclarent avoir l'intention de voter pour C.

a) Si les élections avaient lieu le jour du sondage, C gagnerait-il les élections ? (Proposer un test au niveau de risque de 5%).

b) Le candidat est déçu. Il espérait plus de précision de ce sondage. Combien de personnes ayant l'intention d'exprimer leur suffrage aurait-il fallu interroger pour conclure, au niveau de risque de 5%, que C gagnerait les élections si les élections avaient lieu le jour du sondage ?

Exercice 8-12 : Le diamètre de la prune d'une certaine variété est une variable aléatoire X qu'on suppose normale. Les mesures faites sur un échantillon de 375 prunes de cette variété ont donné les résultats suivants :

diamètre en cm	24	26	28	30	32	34	36	38	40
effectif	7	20	38	79	84	75	53	15	4

a) Estimer, en cm, l'espérance et l'écart-type de X.

b) Quel est l'intervalle de confiance au niveau 0,95 de l'estimation de E(X) ?

IX- Tests portant sur l'égalité des espérances de plusieurs variables aléatoires

1- Egalité des espérances de deux variables normales

Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons aléatoires indépendants, le premier de loi $\mathcal{N}(\mu_1 ; \sigma_1)$, le deuxième de loi $\mathcal{N}(\mu_2 ; \sigma_2)$, où μ_1 et μ_2 sont inconnus.
On suppose qu'on dispose de valeurs expérimentales $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$ des échantillons, qu'on souhaite utiliser pour tester l'hypothèse :

$$(H) : \mu_1 = \mu_2$$

a) variables normales de variances connues

Supposons σ_1 et σ_2 connus.
Construisons le test.

On utilise les estimateurs $\overline{X}$ et $\overline{Y}$ de μ_1 et μ_2, et on note $\overline{x}$ et $\overline{y}$ les estimations expérimentales correspondantes.

On sait que la loi de $(\overline{X} - \overline{Y})$ est normale, d'espérance $(\mu_1 - \mu_2)$ et d'écart-type $\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$. Sous l'hypothèse (H), la variable aléatoire $D = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$ suit donc la loi $\mathcal{N}(0 ; 1)$. Nous la choisissons comme variable discriminante.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\frac{|\overline{x} - \overline{y}|}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où $t_{\alpha/2}$ est défini par :

$$P(Z < -t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha/2 \ , \quad Z \text{ suivant la loi } \mathcal{N}(0 , 1).$$

Exercice 9-1 : Pour mesurer de pH d'une solution, on utilise un pH-mètre qui affiche un résultat dont la loi est $\mathcal{N}(\mu ; 0{,}05)$, où μ est la vraie valeur du pH de la solution. On a mesuré le pH d'une solution A par 12 mesures indépendantes et trouvé une moyenne de 7,04 , et le pH d'une solution B par 10 mesures indépendantes et trouvé une moyenne de 7,05. Peut-on considérer que les deux solutions ont même pH ?

b) variables normales de même variance inconnue

Soit S_1 l'estimateur usuel de σ_1 associé à l'échantillon $X_1, \ldots, X_n$, et notons s_1 l'estimation expérimentale correspondante. Définissons de même S_2 et s_2.

Supposons maintenant les écart-types σ_1 et σ_2 inconnus, mais égaux. Notons σ leur valeur commune.
On peut alors montrer que, sous l'hypothèse (H) ,

$$\frac{(\bar{X} - \bar{Y})}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_1^2 + (m-1)S_2^2}{n + m - 2}}}$$

suit la loi de Student à (n+m-2) degrés de liberté. Nous choisissons cette variable aléatoire comme variable discriminante.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\dfrac{|\bar{x} - \bar{y}|}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)s_1^2 + (m-1)s_2^2}{n + m - 2}}} > t_{\alpha/2}$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où $t_{\alpha/2}$ est défini par :

$P(Z < - t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha /2$, Z suivant la loi de Student t_{n+m-2} .

Exercice 9-2 : Pour mesurer de pH d'une solution, on utilise un nouveau pH-mètre qui affiche un résultat dont la loi est $\mathcal{N}(\mu ; \sigma)$, où μ est la vraie valeur du pH de la solution et où σ n'a pas été déterminé. On a mesuré le pH d'une solution A par 12 mesures indépendantes et trouvé une moyenne de 7,04 et un écart-type empirique de 0,04 , et le pH d'une solution B par 10 mesures indépendantes et trouvé une moyenne de 7,05 et un écart-type empirique de 0,08. Peut-on considérer que les deux solutions ont même pH ?

c) variables normales de variances inconnues

Si les écart-types σ_1 et σ_2 sont inconnus, et si on n'a pas de raison de les présupposer égaux, on ne peut pas travailler comme dans le paragraphe précédent. En effet, la loi de la fonction discriminante qu'on a proposée dépend alors de la valeur des paramètres inconnus σ_1 et σ_2 et ne peut donc plus être utilisée.
Cependant, si les tailles n et m des échantillons sont très grandes, on pourra considérer que les estimations expérimentales s_1 et s_2 des écart-types sont pratiquement égales à leurs vraies valeurs σ_1 et σ_2, et se ramener ainsi au cas du paragraphe a).
Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consistera alors à :

- rejeter l'hypothèse (H) si $\dfrac{|\bar{x} - \bar{y}|}{\sqrt{\dfrac{s_1^2}{n} + \dfrac{s_2^2}{m}}} > t_{\alpha/2}$,

- considérer que l'expérience ne contredit pas (H) sinon ,

où $t_{\alpha/2}$ est défini par :

$P(Z < - t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha /2$, Z suivant la loi $\mathcal{N}(0 , 1)$.

2- Egalité de deux probabilités

Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons aléatoires indépendants, le premier de loi de Bernoulli $\mathcal{B}(p_1)$, le deuxième de loi $\mathcal{B}(p_2)$, où p_1 et p_2 sont inconnus.
On suppose qu'on dispose des échantillons expérimentaux $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_m)$.

Soit à tester l'hypothèse :

$$(H) : p_1 = p_2$$

Supposons que les tailles n et m des échantillons sont grandes.

Pour construire le test, on utilise les estimateurs classiques $\bar{X}$ et $\bar{Y}$ de p_1 et p_2. On note p_1^e et p_2^e leurs valeurs expérimentales.
Supposons (H) vraie et notons p la valeur commune à p_1 et p_2. Alors, d'après le théorème central-limite, $\dfrac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p(1-p)}}$ suit une loi proche de $\mathcal{N}(0 ; 1)$.

Cette variable aléatoire ne peut pas être choisie comme fonction discriminante, car le paramètre p est inconnu. Cependant, comme (H) est vraie, $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ est un échantillon aléatoire de taille n+m de loi $\mathcal{B}(p)$, et on peut l'utiliser pour estimer p. Notons p^e la valeur expérimentale de l'estimateur $\dfrac{X_1 + \ldots + X_n + Y_1 + \ldots + Y_m}{n + m}$ de p.

On considère alors que $\dfrac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p^e(1-p^e)}}$ suit une loi proche de $\mathcal{N}(0 ; 1)$, et c'est cette variable aléatoire qu'on prend fonction discriminante.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste alors à :

- rejeter l'hypothèse (H) si $\dfrac{|p_1^e - p_2^e|}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p^e(1-p^e)}} > t_{\alpha/2}$,

- considérer que l'expérience ne contredit pas (H) sinon .

où $t_{\alpha/2}$ est défini par :

$$P(Z < - t_{\alpha/2}) = P(Z > t_{\alpha/2}) = \alpha /2 \ , \quad Z \text{ suivant la loi } \mathcal{N}(0 , 1).$$

Ce test ne peut être justifié que si les tailles n et m des échantillons sont grandes. On peut dans le cas contraire utiliser un autre test, celui de Fisher-Irwin, qui est basé sur l'expression des probabilités conditionnelles :

$$P \{ X_1 + \ldots + X_n = i \mid X_1 + \ldots + X_n + Y_1 + \ldots + Y_m = k \}.$$

Exercice 9-3 : Pour mesurer le taux d'occupation d'un matériel, on tire au hasard un echantillon d'instants, et en chacun de ces instants, on regarde si le matériel est ou non occupé. On a obtenu les observations suivantes :

	janvier	février
occupation	400	300
inoccupation	100	100
total	500	400

Les taux d'occupation des mois de janvier et février sont-ils significativement différents ?

3- Egalité des espérances de plusieurs variables normales : méthode de la variance

Soient $X_{i1}, \ldots, X_{in_i}$ (i = 1 à m) m échantillons aléatoires indépendants, de lois normales $\mathcal{N}(\mu_i , \sigma)$ d'espérances $\mu_1, \ldots, \mu_m$ inconnues, et de variance inconnue mais commune σ. On notera $n = \sum_{i=1}^{m} n_i$ le nombre total de variables aléatoires.

On suppose qu'on dispose de valeurs expérimentales $x_{i1}, \ldots, x_{in_i}$ (i = 1 à m) de ces échantillons, qu'on souhaite utiliser pour tester l'hypothèse :

$$(H) : \ \mu_1 = \mu_2 = \ldots = \mu_m$$

Pour construire le test de (H), nous allons proposer deux estimateurs de la variance σ^2, le premier convergeant vers σ^2 que l'hypothèse (H) soit ou non vérifiée, le deuxième ne convergeant vers σ^2 que si (H) est vraie, et, dans le cas contraire, surestimant la valeur de σ^2.

Notons $\bar{X}_i$ et S_i^2 les estimateurs usuels de μ_2 et σ^2 associés à l'échantillon $X_{i1}, \ldots, X_{in_i}$:

$$\bar{X}_i = \frac{X_{i1} + \ldots + X_{in_i}}{n_i} \qquad S_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}$$

On pose :

$$S^2_{intra} = \frac{\sum_{i=1}^{m} (n_i - 1)\, S_i^2}{n - m} \qquad \text{(variance intra-classes)}$$

On sait que $(n_i - 1)\frac{S_i^2}{\sigma^2}$ suit la loi $\chi^2_{n_i}$ et que ces m variables aléatoires sont indépendantes.

Par conséquent, $\frac{(n-m)\, S^2_{intra}}{\sigma^2}$ suit la loi du khi-2 à $\sum_{i=1}^{m} (n_i - 1)$ degrés de liberté, c'est-à-dire la loi χ^2_{n-m}. On a donc :

$$E(\ S^2_{intra}\ ) = \sigma^2$$

S^2_{intra} est donc un estimateur sans biais de σ^2, que l'hypothèse (H) soit ou non vérifiée.

Posons maintenant :

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{m} n_i\ \bar{X}_i}{\sum_{i=1}^{m} n_i} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{m} n_i} \qquad \text{(moyenne globale)}$$

$$S^2_{inter} = \frac{\sum_{i=1}^{m} n_i\ (\ \bar{X}_i - \bar{\bar{X}}\ )^2}{m - 1} \qquad \text{(variance inter-classe)}$$

On peut montrer par un calcul que :

$$E(\ S^2_{inter}\ ) = \sigma^2 + \frac{1}{m-1} \sum_{i=1}^{m} n_i\ (\ \mu_i - \bar{\mu}\ )^2 \qquad \text{où} \qquad \bar{\mu} = \frac{\sum_{i=1}^{m} n_i \mu_i}{\sum_{i=1}^{m} n_i}$$

Ainsi, si l'hypothèse (H) est fausse, S^2_{inter} n'est pas un estimateur de σ^2, il en surestime sa valeur.

Supposons maintenant (H) vraie. Alors, S^2_{inter} est un estimateur sans biais de σ^2. On peut aussi montrer, mais la preuve n'en est pas élémentaire, que $\frac{(m-1)\, S^2_{inter}}{\sigma^2}$ suit la loi χ^2_{m-1} et que les variables aléatoires S^2_{inter} et S^2_{intra} sont indépendantes. On en conclut que la variable aléatoire $\frac{S^2_{inter}}{S^2_{intra}}$ suit la loi $F_{m-1,\ n-m}$. C'est cette variable qu'on choisit comme variable discriminante. Notons $\frac{s^2_{inter}}{s^2_{intra}}$ sa valeur expérimentale.

Le test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste à :

- rejeter l'hypothèse (H) si $\frac{s^2_{inter}}{s^2_{intra}} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où t_α est défini par :

$$P(Z > t_\alpha) = \alpha \ , \quad Z \text{ suivant la loi } F_{m-1,\ n-m}.$$

Remarque : Dans la pratique, il est souhaitable que les tailles n_i des échantillons soient égales ou presque. Dans ce cas en effet, d'une part on risque moins de considérer comme acceptable l'hypothèse (H) alors qu'elle est fausse, d'autre part, le test est encore relativement bon si les variances des m échantillons ne sont pas tout à fait égales.

Exercice 9-4 : Pour comparer trois types d'essence, on a mesuré la consommation d'essence à vitesse stabilisée de 90km/h de 18 voitures à peu près identiques et obtenu le tableau suivant, où les données sont exprimées en nombre de litres pour 100 km :

essence 1	5,50	6,3	5,95	6,15	6,5	5,6
essence 2	6,1	5,9	6,45	6,05	5,52	5,75
essence 3	6,35	6,8	5,8	5,95	6,4	6,25

La consommation de ces voitures dépend-elle du type d'essence utilisé ?

Exercice 9-5 : Reprendre les données de l'exercice 9-2 avec méthode de la variance. Comparer les conclusions obtenues avec les deux méthodes.

4- Exercices

Exercice 9-6 : On souhaite étudier les effets secondaires d'un certain médicament sur le rythme cardiaque. Pour cela, on a pris le pouls de 11 personnes avant et après la prise de ce médicament, et obtenu les résultats suivants, exprimés en nombre de pulsations par minute :

patient	1	2	3	4	5	6	7	8	9	10	11
avant	74	86	62	98	102	78	64	84	68	79	70
après	70	85	63	90	110	71	60	80	67	69	74

Proposer un test adapté à ces données, en précisant ce qu'il faut supposer pour le justifier. Le mettre en œuvre.

Exercice 9-7 : a) On dispose des notes obtenues à un devoir surveillé par les 24 et 25 étudiants de deux groupes de TD. Quel test proposer pour comparer le niveau de réussite des deux groupes ? Que doit-on supposer pour le justifier ?

b) 10 copies d'examen ont été corrigées par deux correcteurs A et B. Pour chaque copie, on connaît la note donnée par A et la note donnée par B. Quel test proposer pour comparer la sévérité des correcteurs ? Que doit-on supposer pour le justifier ?

Exercice 9-8 : Un constructeur A affirme que la charge de rupture de ses câbles est plus grande que celle des câbles du constructeur B. Pour s'en assurer, un client a fait mesuré la charge de rupture de 14 câbles et trouvé :

câbles A	140	138	143	142	144	137	141	139
câbles B	135	140	136	142	138	140		

Tester l'affirmation du constructeur au risque 0,05.

Exercice 9-9 : Un laboratoire pharmaceutique peut fabriquer un même médicament suivant deux procédés différents, équivalents du point de vue de leur coût. On a mesuré la durée de conservation du médicament par 20 expériences indépendantes et obtenu les durées suivantes, exprimées en nombre d'années :

Procédé A	2,5	3	2	1,5	3,5	1	4	4,5	0,5	2,5
Procédé B	2,2	2,3	2,5	2,8	2,7	2,3	2,8	2,5	2	2,9

Pour chacun des procédés, quels sont les moyenne et écart-type empiriques des résultats? A votre avis, l'un des procédés est-il préférable ?

Exercice 9-10 : Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons aléatoires indépendants, le premier de loi $\mathcal{N}(\mu_1 ; \sigma_1)$, le deuxième de loi $\mathcal{N}(\mu_2 ; \sigma_2)$, où μ_1, μ_2, σ_1 et σ_2 sont inconnus. Proposer un test de l'hypothèse ($\sigma_1 = \sigma_2$). Le mettre en œuvre avec les données de l'exercice 9-2.

X- Tests d'hypothèses non-paramétriques sur la loi d'une variable aléatoire

Exemple 10-1 : On a lancé un dé 360 fois et obtenu le tableau :

n° de la face	1	2	3	4	5	6
effectif	43	55	51	71	72	68

Comment utiliser ces données pour tester l'hypothèse que toutes les faces ont la même probabilité ? •

Exemple 10-2 : Dans les exemples des deux derniers chapitres, nous avons souvent supposé que la loi d'un échantillon dont on disposait d'une valeur expérimentale suivait une loi normale. Comment tester une telle hypothèse ? Supposons par exemple que Z_1, …, Z_n est un échantillon aléatoire de loi inconnue, et que nous voulons tester l'hypothèse :

$$(H) : \text{ la loi de } Z_1, \ldots, Z_n \text{ est } \mathcal{N}(0 , 1).$$

à l'aide de valeurs expérimentales z_1, …, z_n.

Contrairement au cas précédent, la loi de référence est ici continue. On se ramène au cas discret en découpant l'ensemble des valeurs possibles des variables aléatoires Z_1, …, Z_n en un nombre fini k de régions, en général des intervalles, R_1, …, R_k. Les données expérimentales z_1, …, z_n se répartissent suivant le tableau d'effectifs :

région	R_1	R_2	……	R_k
effectif	c_1	c_2	……	c_k

Si l'hypothèse (H) est vraie, on sait calculer la probabilité p_a pour qu'un résultat Z tombe dans la zone R_a. On est donc ramené à une situation analogue à celle de l'exemple 10-1. Il faudra cependant être plus prudent dans l'interprétation du résultat du test, car il peut dépendre de la manière dont les régions R_a ont été délimitées. •

1- Egalité de la loi de l'échantillon et d'une loi spécifiée

Soit Y_1, …, Y_n un échantillon aléatoire à valeur dans {1, 2, …, k} de loi inconnue. Pour simplifier la présentation, notons Y une variable aléatoire de même loi. Nous supposons disposer d'une valeur expérimentale y_1, …, y_n de l'échantillon, et nous voulons l'utiliser pour tester l'hypothèse :

$$(H) : \forall\ a \in \{1, 2, \ldots, k\} \quad P\{ Y = a \} = p_a$$

où les probabilités p_a sont données et vérifient $\sum_{a=1}^{k} p_a = 1$.

Pour a dans {1, 2, …, k}, posons :

$$C_a = \text{card } \{ i \ / \ Y_i = a \}$$

et notons c_a la valeur expérimentale correspondante.

Sous l'hypothèse (H), C_a est une variable aléatoire de loi $\mathfrak{B}(n , p_a)$. Son espérance est np_a . La valeur prise par $(C_a - np_a)^2$, lorsque n est grand, donne donc une indication de la plausibilité de l'hypothèse que C_a est une variable aléatoire de loi $\mathfrak{B}(n , p_a)$: plus cette valeur est grande, moins cette hypothèse est plausible.
De fait, on choisit comme fonction discriminante :

$$D = \sum_{a=1}^{k} \frac{(C_a - np_a)^2}{np_a}$$

et on décidera de rejeter (H) lorsque la valeur expérimentale d de D est trop grande.

Remarque : Dans le contexte d'utilisation de ce test, la valeur prise par Y n'intervient que comme un outil pour classer les individus de la population étudiée. La fonction discriminante D est définie à partir des contingents des différentes classes de l'échantillon observé. Y pourrait tout autant être une variable aléatoire qualitative, comme dans l'exemple 10-2, au lieu d'être numérique. •

a) Test du khi-deux

On peut montrer, - mais, dès que k est plus grand que 2, la preuve n'est pas élémentaire -, que si (H) est vraie et si n est grand, D suit approximativement la loi χ^2_{k-1}.
Dans la pratique, on utilise cette approximation si pour tout a, $np_a \geq 1$ et si pour au moins 80% des a, $np_a \geq 5$.

La mise en œuvre du test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\sum_{a=1}^{k} \frac{(c_a - np_a)^2}{np_a} > t_\alpha$,
- considérer que l'expérience ne contredit pas (H) sinon ,

où t_α est défini par :

$$P(Z > t_\alpha) = \alpha \ , \quad Z \text{ suivant la loi } \chi^2_{k-1}$$

Exercice 10-1 : a) Les données sont celles de l'exemple 10-1. Tester l'hypothèse que toutes les faces ont la même probabilité, au niveau de risque de 2%, puis au niveau de risque de 5%.
b) Supposer toutes les effectifs multipliés par 2, et tester la même hypothèse.

b) Test par simulation

Notons encore d la valeur expérimentale de D.

Si la taille de l'échantillon ne permet pas l'approximation de la loi de la fonction discriminante D par une loi du khi-deux, on peut utiliser une simulation de cette loi sur ordinateur :

- On tire indépendamment n valeurs y suivant la loi de Y donnée par l'hypothèse (H), et on calcule la valeur de D correspondante. Notons-la d_1.
- On recommence un grand nombre r de fois ce tirage. On obtient les valeurs $d_1, \ldots, d_r$.
- De la loi des grands nombres, on déduit que, sous l'hypothèse (H) :

$$P\{ D \geq d \} \approx \frac{\text{card}\{ i \ / \ d_i \geq d \}}{r}$$

La mise en œuvre du test de (H) au niveau de risque α $(0 < \alpha < 1)$ consiste donc à :

- rejeter l'hypothèse (H) si $\frac{\text{card}\{ i \ / \ d_i \geq d \}}{r} < \alpha$,
- considérer que l'expérience ne contredit pas (H) sinon .

2- Cas où certains paramètres ne sont pas spécifiés

Exemple 10-3 : Reprenons l'exemple 10-2, mais supposons maintenant à tester l'hypothèse :

(H) : la loi de $Z_1, \ldots, Z_n$ est normale.

Sous cette seule hypothèse, les probabilités p_a pour qu'un résultat Z tombe dans la zone R_a ne sont pas calculables. Pour tester (H_Z), on estime les paramètres μ et σ de la loi de l'échantillon $Z_1, \ldots, Z_n$ par les estimateurs usuels $\overline{X}$ et S. On teste ensuite l'hypothèse :

(H') : la loi de $Z_1, \ldots, Z_n$ est $\mathcal{N}(\overline{x}, s)$

comme dans le paragraphe précédent, soit par simulation, soit par le test du khi-deux, le nombre de degrés de liberté étant alors (k - 1- e), où e est le nombre de paramètres estimés (ici, e=2).•

Exercice 10-2 : On a relevé le nombre d'accidents durant une période de 30 semaines dans un secteur donné, et obtenu :

8	0	0	1	3	4	0	2	12	5
1	8	0	2	0	1	9	3	4	5
3	3	4	7	4	0	1	2	1	2

Peut-on considérer que ce nombre suit une loi de Poisson ? (On utilisera la partition de l'ensemble des valeurs possibles : {0} {1} {2, 3} {4, 5} {6 ou plus}).

3- Egalité des lois de plusieurs échantillons

Soient $Y_{i1}, \ldots, Y_{in_i}$ (i = 1 à m) m échantillons aléatoires indépendants de lois inconnues, toutes les variables aléatoires prenant leurs valeurs dans {1, 2, …, k}, et soit à tester l'hypothèse (H) :

(H) : Les lois des m échantillons sont identiques

Notons $y_{i1}, \ldots, y_{in_i}$ les valeurs expérimentales des échantillons.

Pour simplifier la présentation, notons, pour tout i, Y_i une variable aléatoire ayant la même loi que l'échantillon $Y_{i1}, \ldots, Y_{in_i}$. Avec cette notation, (H) se réécrit :

$$(H) : \quad \forall\ a \in \{1, 2, \ldots, k\} \quad P\{ Y_1 = a \} = \ldots = P\{ Y_m = a \}$$

Supposons d'abord l'hypothèse (H) vraie. Notons alors Y une variable aléatoire ayant la même loi que les Y_i. Pour a dans {1, 2, …, k}, estimons les probabilités P{ Y = a } par :

$$p_a = \frac{\text{card}\{ (i,j) \ / \ y_{ij} = a \}}{\sum_{i=1}^{m} n_i}$$

Posons :

$$(H') \ : \forall\ a \in \{1, 2, \ldots, k\} \quad \forall\ i \in \{1, 2, \ldots, m\} \quad P\{ Y_i = a\} = p_a$$

On teste (H') par une méthode semblable à celle du paragraphe 1. On définit pour cela :

$$D = \sum_{i=1}^{m} \sum_{a=1}^{k} \frac{(C_{ia} - n_i p_a)^2}{n_i p_a}$$

où :

$$C_{ia} = \text{card}\{ j \ / \ Y_{ij} = a \}$$

et on décide de rejeter (H), lorsque la valeur expérimentale d de D est trop grande.

On procède soit par simulation, soit par le test du khi-deux, le nombre de degrés de liberté étant alors (k-1)(m-1).

Exercice 10-3 : On a testé trois modèles de machines à laver, A, B et C, en comptant le nombre de pannes durant leur 3 premières années de fonctionnement. On a obtenu le tableau :

	0 panne	1 panne	2 pannes	3 pannes ou plus
A	884	403	95	23
B	123	693	373	28
C	57	219	144	8

Cette expérience met-elle en évidence une différence entre les trois modèles ?

4- Indépendance de deux caractères aléatoires

On étudie conjointement deux caractères des individus d'une population, qui prennent leurs valeurs respectivement dans {1, 2, …, k} et {1, 2, …, m}. On suppose disposer de n valeurs expérimentales indépendantes $(x_1, y_1), \ldots, (x_n, y_n)$.

Pour représenter cette situation, on introduit (X_1,Y_1), …, (X_n,Y_n) un échantillon de n variables aléatoires indépendantes à valeurs dans $\{1, 2, \ldots, k\}\times\{1, 2, \ldots, m\}$ de même loi (inconnue). Notons (X,Y) une variable aléatoire de même loi. Quels que soient les couples (a_i, b_i) on a donc, par hypothèse :

$$P\{ [(X_1,Y_1) = (a_1, b_1)] \text{ et } \ldots \text{ et } [(X_n,Y_n) = (a_n, b_n)] \} =$$
$$= P\{ (X_1,Y_1) = (a_1, b_1) \} \ldots P\{ (X_n,Y_n) = (a_n, b_n) \} =$$
$$= P\{ (X,Y) = (a_1, b_1) \} \ldots P\{ (X,Y) = (a_n, b_n) \}$$

Soit à tester l'hypothèse d'indépendance des caractères, autrement dit l'hypothèse :

$(H) : \forall (a, b)\in \{1, 2, \ldots, k\}\times\{1, 2, \ldots, m\}\ P\{ (X,Y) = (a, b) \} = P\{ X = a \}P\{ Y = b \}$

On estime les lois (marginales) de X et Y par :

$$p_a^X = \frac{card\{ i \ / \ x_i = a \}}{n} \qquad a \in \{1, 2, \ldots, k\}$$

$$p_b^Y = \frac{card\{ i \ / \ y_i = b \}}{n} \qquad b \in \{1, 2, \ldots, m\}$$

On choisit comme fonction discriminante :

$$D = \sum_{a=1}^{k} \sum_{b=1}^{m} \frac{(C_{(a,b)} - np_a^X p_b^Y)^2}{np_a^X p_b^Y}$$

où :

$$C_{(a,b)} = card\{ i \ / \ (X_i, Y_i) = (a, b) \}$$

et on décide de rejeter (H) lorsque la valeur expérimentale d de D est trop grande.

On procède soit par simulation, soit par le test du khi-deux, le nombre de degrés de liberté étant alors (k-1)(m-1).

Exercice 10-4 : On a interrogé 2000 personnes lors de leur départ en vacances sur leur destination et le moyen de transport utilisé pour s'y rendre. On a obtenu le tableau :

	Campagne	Mer	Montagne
Voiture	250	700	350
Train	200	200	50
Avion	15	200	35

Y a-t-il un lien entre la destination et le moyen de transport ?

5- Test des signes

Exemple 10-4 : On a testé un médicament contre l'hypertension sur 18 patients en mesurant la différence entre leur tension avant le début du traitement et après un mois de traitement. On a obtenu les résultats :

-2	-1	+1	+3	-8	+1	+2	-4	-5
-3	-3	-6	-2	-7	+2	-7	-5	-4

On se demande si le médicament a un effet réel sur l'hypertension, - ou s'il est efficace contre l'hypertension -.
Notons X la variable aléatoire qui représente cette différence.
- Si on peut supposer que la loi de X est normale, on peut utiliser un test de Student de l'hypothèse simple "l'espérance de X est nulle" ou de l'hypothèse composite "l'espérance de X est négative". (Le fait que la loi de X peut être considéré comme normale peut lui-même être testé par un test du khi-deux, mais le test sera ici grossier car l'effectif total est faible.)
- Si on ne peut pas supposer la loi normale, on peut proposer de tester une l'hypothèse sur la valeur de sa médiane. •

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de loi inconnue, de médiane m. On note F sa fonction de répartition, qu'on suppose pour simplifier continue.
Soit à tester l'hypothèse :

$$(H) : \quad m = m_0$$

où m_0 est un réel spécifié.
Introduisons les variables aléatoires (indépendantes) Y_i :

$$Y_i = 1 \quad \text{si} \quad X_i \leq m_0 ,$$
$$= 0 \quad \text{sinon} .$$

Elles suivent la loi de Bernoulli $\mathcal{B}(F(m_0))$. L'hypothèse (H) équivaut donc à l'hypothèse :

$$(H') : \text{ le paramètre de la loi de Bernoulli des } Y_i \text{ vaut } \frac{1}{2} .$$

On est ramené au cas traité dans paragraphe 4 du chapitre VIII.
On pose donc :

$$D = \text{card}\{ i \ / \ X_i \leq m_0 \}$$

et on note d sa valeur expérimentale.

Le test consiste à :
- rejeter (H) si $\alpha > 2\, P\{ Z < \min(d, n-d) \}$,
- considérer que l'expérience ne contredit pas (H) sinon,

où Z suit la loi $\mathcal{B}(n , \frac{1}{2})$.

Exercice 10-5 : a) Utiliser le test des signes pour traiter l'exemple 10-4.
b) Remarquer qu'on peut aussi tester l'hypothèse (H') du test des signes par un test du khi-deux. Que trouve-t-on par cette méthode ?
c) Que conclut-on en utilisant un test de Student ?

6- Exercices

Exercice 10-6 : Reprendre les données de l'exercice 8-12, et tester la normalité de la loi de X.

Exercice 10-7 : Proposer une deuxième façon de traiter l'exercice 9-3.

Exercice 10-8 : Sur 100 tubes à vide testés, 41 ont eu une durée de vie de moins de 30 heures, 31 entre 30 et 60 heures, 13 entre 60 et 90 heures, et 15 plus de 90 heures. Ces données sont-elles compatibles avec l'hypothèse que la durée de vie d'un tube à vide est une loi exponentielle d'espérance égale à 50 heures ?

Exercice 10-9 : Le tableau ci-dessous donne la répartition de 200 naissances en fonction de la parité de la mère et du poids du nouveau-né.

	primipares	multipares
poids inférieur à 3kg	26	20
entre 3 et 4 kg	61	63
supérieur à 4 kg	8	22

Les deux caractères, parité de la mère et poids du nouveau-né, sont-ils statistiquement reliés ?

Exercice 10-10 : 2000 personnes ont passé un concours. Proposer une méthode de comparaison des manières de noter de deux correcteurs A et B, sachant qu'on peut pour cela demander à chacun de corriger 50 copies.

Exercice 10-11 : Reprendre l'exercice 9-6 en utilisant le test des signes.