1 Exercices

Rappels.

A. La loi du χ^2 à p degrés de liberté est la loi de $\sum_{i=1}^{p} X_i^2$ lorsque les v.a. $(X_i)_{i=1}^{p}$ sont i.i.d. de loi $\mathcal{N}(0,1)$.

B. Soit W une variable gaussienne standardisée et V une variable du χ^2 centrée à p degrés de liberté, indépendante de W. Alors la variable

$$T = \frac{W+\gamma}{\sqrt{V/p}}$$

est distribuée suivant une loi de Student décentrée, à p degrés de liberté et de paramètre de non-centralité γ.

C. Si U et V sont deux v.a. indépendantes de loi resp. $\chi^2(n)$ et $\chi^2(p)$ alors

$$\frac{U/n}{V/p}$$

suit une loi de Fisher de paramètres (n,p).

Exercice 1. Soit $(X_1,\dots,X_n)$, un n-échantillon de la loi gaussienne $\mathrm{N}(\mu,\sigma^2)$ où $\mu \in \mathbb{R}$ est un paramètre inconnu et la variance σ^2 est connue. Soit $\mu_0 \in \mathbb{R}$. On considère le test

$$H_0 :\ \mu = \mu_0, \qquad \text{contre} \qquad H_1 :\ \mu \neq \mu_0$$

On pour $c > 0$, on considère le test pur $\delta_c(X_1,\dots,X_n) = \mathbb{1}_{\{|T(X_1,\dots,X_n)|\geq c\}}$ où la statistique de test $(X_1,\dots,X_n)$ est donnée par :

$$T(X_1,\dots,X_n) := \left|\sqrt{n}\frac{(\bar{X}_n-\mu_0)}{\sigma}\right|, \qquad \text{où} \quad \bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i\ .$$

1. Déterminer en fonction du seuil critique c la fonction puissance du test $\mu \mapsto \beta_c(\mu)$.
2. Déterminer en fonction du seuil critique c la taille du test (ou risque de première espèce).
3. Calculer la p-valeur associée à l'observation Z.

On suppose maintenant que la variance σ^2 est inconnue et on pose

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\ .$$

4. Montrer que les statistiques $\bar{X}_n$ et S_n^2 sont indépendantes.
5. Montrer que $(n-1)S_n^2/\sigma^2$ est distribuée suivant une loi du χ^2 à $(n-1)$ degrés de liberté.

	Échantillon #1	Échantillon #2
Nombre d'observations	120	120
Moyenne	75	74.99
Ecart-type	8.72	6.54

6. Soit $\mu_0 \in \mathbb{R}$. Déterminer la distribution de

$$\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} .$$

On considère le test $H_0 : \ \mu = \mu_0, \qquad$ contre $\qquad H_1 : \ \mu \neq \mu_0$. On utilise la statistique de test $T(Z) := |\sqrt{n}(\bar{X}_n - \mu_0)/S_n|$.

7. Déterminer en fonction du seuil critique c, la fonction puissance du test.
8. Déterminer en fonction du seuil critique c la taille du test (ou risque de première espèce).
9. Déterminer le seuil critique de façon à obtenir un test de niveau $\alpha \in]0,1[$.
10. Calculer la p-valeur de l'observation Z.

Exercice 2 (Modèle gaussien à deux échantillons). On dispose d'une variable qualitative binaire qui permet de définir deux groupes.

— On mesure une variable quantitative qui permet de calculer dans chaque groupe les différents paramètres de la distribution : moyenne, estimateur de l'écart type...

— On désire savoir si les moyennes observées dans chacun des groupes peuvent être considérées comme des estimateurs de la même moyenne aux fluctuations du hasard près [1].

Nous formalisons cette hypothèse de test ci-dessous. Soient $X_1, \dots, X_n$ un échantillon i.i.d. de $\mathrm{N}(\mu_0, \sigma^2)$ et $Y_1, \dots, Y_p$ un échantillon i.i.d. de $\mathrm{N}(\mu_1, \sigma^2)$ où $\theta := (\mu_0, \mu_1, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^*$. On suppose de plus que les vecteurs aléatoires $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_p)$ sont indépendants.

1. Déterminer l'estimateur du maximum de vraisemblance du paramètre θ.
2. Construire un intervalle de confiance de $\mu_0 - \mu_1$ de niveau de confiance $1 - \alpha$.
3. En déduire un test de niveau α de l'hypothèse

$$H_0 : \ \mu_0 = \mu_1, \qquad \text{contre} \qquad H_1 : \ \mu_0 \neq \mu_1 .$$

4. Construire un test de niveau α de l'hypothèse

$$H_0 : \ \mu_0 \leq \mu_1, \qquad \text{contre} \qquad H_1 : \ \mu_0 > \mu_1.$$

On utilise maintenant ce test pour comparer la consistance de deux correcteurs de l'examen d'entrée à l'Ecole Polytechnique. Nous avons visualisé les notes pour les deux correcteurs (sur 100) dans le tableau suivant

1. vous pourrez consulter `https://www.youtube.com/watch?v=oJjkjkY6mmA`

5. La p-valeur du test est de 0.98. Que concluez vous ?

Les v.a. $(X_1,\cdots,X_n,Y_1,\cdots,Y_p)$ restent indépendantes, mais on suppose maintenant que $X_1,\ldots,X_n$ est un échantillon i.i.d. de loi $\mathrm{N}(\mu_0,\sigma_0^2)$ et $Y_1,\ldots,Y_p$ est un échantillon i.i.d. de loi $\mathrm{N}(\mu_1,\sigma_1^2)$ où $\theta := (\mu_0,\mu_1,\sigma_0^2,\sigma_1^2) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}_+^*\times\mathbb{R}_+^*$.

6. Construire une fonction pivotale du paramètre σ_1^2/σ_0^2.
7. En déduire un test de niveau α de l'hypothèse

$$H_0:\ \sigma_0^2 \le \sigma_1^2, \qquad \text{contre} \qquad H_1:\ \sigma_0^2 > \sigma_1^2\ .$$

8. Calculer la p-valeur de ce test.

Exercice 3. Des plaignants [2] ont poursuivi en justice le Ministère de la Santé suite à une campagne de vaccination menée sur des enfants et ayant entraîné des dommages fonctionnels irréversibles pour certains d'entre eux. Ce vaccin était connu pour entraîner ce type de dommages en de très rares circonstances. Des études antérieures menées dans d'autres pays ont montré que ce risque était en moyenne d'un cas sur 310 000 vaccinations. Les plaignants avaient été informés de ce risque et l'avaient accepté. Les doses de vaccin ayant provoqué les dommages objet de la plainte provenaient d'un lot ayant servi à vacciner un groupe de 300 533 enfants. Dans ce groupe, quatre cas de dommages ont été détectés.

1. On modélise l'événement "le vaccin provoque des dommages fonctionnels irréversibles sur l'enfant i" par une variable aléatoire de Bernoulli, X_i, de paramètre p. Calculer la valeur p_0 correspondant aux résultats des études antérieures.

On considère le test

$$H_0:\ p \le p_0, \qquad \text{contre} \qquad H_1:\ p > p_0$$

L'hypothèse nulle correspond au risque que les clients avaient accepté.

2. On pose $S_n = \sum_{i=1}^n X_i$ avec $n = 300533$. Pour $\alpha \in [0,1]$, déterminer un test de niveau α de basé sur la statistique de test S_n.
3. On a $\mathbb{P}_{p_0}(S_n > 0) = 0.6207$, $\mathbb{P}_{p_0}(S_n > 1) = 0.2530$, $\mathbb{P}_{p_0}(S_n > 2) = 0.0748$, $\mathbb{P}_{p_0}(S_n > 3) = 0.0172$, $\mathbb{P}_{p_0}(S_n > 4) = 0.0032$. L'hypothèse est elle acceptée au niveau 0.05 ? au niveau 0.01 ?
4. Quelle est la p-valeur de l'observation ?

On peut modéliser la loi du nombre S_n de cas de dommages par une loi de Poisson de paramètre θ. [3]

5. Calculer la valeur θ_0 attendue si le vaccin est conforme aux études antérieures.
6. Les plaignants militent pour que le test prouve effectivement que le nombre moyen d'accidents reste inférieur à θ_0. Pour eux, le nombre moyen est supérieur à θ_0 sauf preuve du contraire. Proposer un test à partir de la variable S_n.

2. M. Aitkin. Evidence and the Posterior Bayes Factor. *Math. Scientist*, vol. 17, pp. 15-25 (1992).
3. regarder `https://www.youtube.com/watch?v=Lygy1OxUG88`

7. Donner la p-valeur de ce test. Accepte-t-on H_0 au seuil de 0.05 ?

Exercice 4 (Intervalle de confiance pour la loi uniforme). Soit $(X_1,\dots,X_n)$, n variables aléatoires réelles, indépendantes, distribuées suivant une loi uniforme sur l'intervalle $[0,\theta]$ où $\theta\in\Theta:=\mathbb{R}_+^*$. On pose $X_{n:n}:=\max(X_1,\dots,X_n)$.

1. Montrer que $X_{n:n}/\theta$ est une fonction pivotale pour θ.
2. Déterminer l'intervalle de confiance de probabilité de couverture $(1-\alpha)$ de longueur minimale.

Exercice 5. Une firme pharmaceutique a mis au point une nouvelle molécule pour faire chuter le taux de sucre dans le sang. Pour tester l'efficacité de cette molécule, elle le compare à un placebo. Elle réunit $n+m$ patients. A un premier groupe de m individus, elle administre un placebo (sans leur dire !). Au second groupe elle donne sa nouvelle molécule. Après un délai approprié, on mesure les taux de glycémie $\{X_i:\ i=1,\ \dots,\ n\}$ et $\{Y_j:\ j=1,\dots,m\}$ chez les deux groupes. [4]

On suppose que les variables aléatoires $(X_1,\dots,X_n,Y_1,\ \dots,\ Y_m)$ sont indépendantes ; que les v.a. $(X_i)_i$ ont même loi de fonction de répartition F_X ; et que les v.a. $(Y_i)_i$ ont même loi de fonction de répartition F_Y. On supposera que les fonctions F_X et F_Y sont continues.

On veut tester si les lois des X et des Y sont les mêmes ou si les Y_i sont stochastiquement plus petits que les X_i, c'est à dire $F_X<F_Y$. On va donc tester $H_0:F_X=F_Y$ contre $H_1:F_X<F_Y$.

On pose $Z_i=X_i$ pour $i=1,\dots,n$ et $Z_{n+i}=Y_i$ pour $i=1,\dots,m$. On note $R(i)$ le rang de Z_i dans la suite $(Z_1,\ \dots,\ Z_{n+m})$, à savoir, $R_i=\ell$ si $Z_i=Z_{\ell:n+m}$. La statistique de Wilcoxon est définie par

$$W_{n,m}:=\sum_{i=1}^{n}R_i\ .$$

L'idée est que sous H_1 la statistique $W_{n,m}$ sera plus grande que sous H_0.

1. Soit $N=n+m$. Montrez que la *statistique de rang* $R:=(R_1,\ \dots,\ R_N)$ suit sous H_0 la loi uniforme sur l'ensemble $\mathscr{S}_N$ de toutes les permutations de $\{1,\dots,N\}$. En déduire que sous H_0 la statistique de Wilcoxon est *libre*, i.e., la loi de $W_{n,m}$ ne dépend pas de F_X. Quelle est la loi de R_i sous H_0 ?
2. Montrez que sous H_0 on a $\mathrm{E}(W_{n,m})=n(n+m+1)/2$.
3. Montrez que sous H_0 on a

$$\begin{aligned}\mathrm{Var}(W_{n,m})&=n\,\mathrm{Var}(R_1)+n(n-1)\,\mathrm{Cov}(R_1,\ R_2)\\ 0=\mathrm{Var}\left(\sum_{i=1}^{n+m}R_i\right)&=(n+m)\mathrm{Var}(R_1)+(n+m)(n+m-1)\mathrm{Cov}(R_1,\ R_2)\,.\end{aligned}$$

   En déduire que, sous H_0, $\mathrm{Var}(W_{n,m})=nm(n+m+1)/12$.
4. Proposer un test de niveau α.

4. voir `https://www.youtube.com/watch?v=QDUQscWyXqw`