UNIVERSITÉ DE RENNES 1

Master STS mention mathématiques

Statistique mathématique

Recueil d'exercices

UFR Mathématiques

Année universitaire 2011–2012

TABULATION DE LA FONCTION DE RÉPARTITION DE LA LOI GAUSSIENNE

c	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Tabulation de la fonction de répartition de la loi normale centrée réduite $\Phi(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} \exp(-t^2/2)dt$. Dans l'exemple ci-dessus, à l'intersection de la ligne grisée qui passe par 2.1 et la colonne griséé qui passe par 0.04, le paramètre c vaut 2.14 et $\Phi(c) = 0.9838$.

Table des matières

1 Quelques résultats sur les lois usuelles en vue de la statistique

Exercice 1 [Loi géométrique] Soit $(E_n)_{n\in\mathbb{N}^*}$ une suite d'événements indépendants définis sur le même espace de probabilité et de même probabilité p, non nulle ; on désigne par X la variable aléatoire égale à l'indice du premier événement qui se réalise.

1. Déterminer la loi de probabilité de la variable aléatoire X, sa moyenne et sa variance.
2. Calculer la probabilité conditionnelle pour que X soit égale à 2, sachant que X est pair ; généraliser ce résultat au cas où X est divisible par un entier k.
3. Pour des entiers positifs k et n, calculer
$$\mathbb{P}(X = k+n|X > k).$$
4. Soit Y la variable aléatoire égale à l'indice n pour lequel, pour la première fois, les événements E_{n-1} et E_n se réalisent ; déterminer la fonction génératrice de Y, en déduire sa loi, sa moyenne et sa variance.

Exercice 2 [Loi normale] Soient $m \in \mathbb{R}$ et $\sigma > 0$. On note
$$g_{m,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-m)^2}{2\sigma^2}),\ \text{pour } x \in \mathbb{R}.$$

1. Montrer que g_{m,σ^2} est une densité de probabilité. (On note la loi correspondante $\mathscr{N}(m,\sigma^2)$ et on l'appelle loi normale de moyenne m et variance σ^2.)
2. Calculer $\max_{x\in\mathbb{R}} g_{m,\sigma^2}(x)$ et $\lim_{|x|\to\infty} g_{m,\sigma^2}(x)$. Tracer sur le même dessin $g_{0,1/4}$, $g_{0,1}$ et $g_{0,4}$.
3. Si X est une variable aléatoire de loi ayant comme densité g_{m,σ^2}, calculer la fonction génératrice des moments $\phi(r) := \mathbb{E}\exp(-rX)$ pour $r \in \mathbb{R}$. Calculer $\mathbb{E}X$ et $\mathsf{Var}X$.
4. La fonction de répartition de la variable aléatoire gaussienne centrée réduite Z, de densité $g_{0,1}$, est tabulée et notée $\Phi(z) = \mathbb{P}(Z \le z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp(-x^2/2)dx$. Soit X une variable aléatoire distribuée selon $\mathscr{N}(m,\sigma^2)$. Calculer à l'aide des tables $\mathbb{P}(1 \le X \le 8)$ pour $m = 5$ et $\sigma = 2$. Trouver t_α tel que $\mathbb{P}(|Z| > t_\alpha) = \alpha$ pour $\alpha = 0.1$.
5. Soient X distribuée selon $\mathscr{N}(m,\sigma^2)$ et Z distribuée selon $\mathscr{N}(0,1)$. Exprimer en termes de Z la probabilité $p_k = \mathbb{P}(|X-m| \le k\sigma)$ et donner les valeurs de p_k pour $k = 1,2,3$.
6. Calculer la fonction caractéristique χ de la loi $\mathscr{N}(m,\sigma^2)$.
7. Soient Z_1 et Z_2 deux variables indépendantes distribuées selon $\mathscr{N}(0,1)$. Montrer que le rapport Z_1/Z_2 est distribué selon la loi de Cauchy.
8. Soient X distribuée selon $\mathscr{N}(m,\sigma^2)$ et $Y = \exp(X)$. Trouver la densité de Y. On appelle la loi de Y log-normale et on note $\mathsf{lognorm}(m,\sigma^2)$.
9. Soit Y une variable aléatoire distribuée selon la loi $\mathsf{lognorm}(m,\sigma^2)$; calculer $\mathbb{E}Y$.
10. Soit Y une variable aléatoire distribuée selon la loi $\mathsf{lognorm}(m,\sigma^2)$; determiner la loi de la variable aléatoire $W = 1/Y$.

Exercice 3 [Loi gamma]

Soient p et λ des réels strictement positifs ; on note $\gamma_{p,\lambda}(x) = C\lambda^p x^{p-1}\exp(-\lambda x)\mathbb{1}_{[0,\infty[}(x)$. Déterminer la constante C (en fonction de p et λ) pour que $\gamma_{p,\lambda}$ soit une densité de probabilité. Dans la suite, la constante C sera fixée à la valeur ainsi déterminée.

1. Tracer sur le même graphique $\gamma_{1,1}, \gamma_{2,1}, \gamma_{4,1}$ et sur le même graphique $\gamma_{2,2}, \gamma_{2,1}, \gamma_{2,1/1}$.
2. Calculer la fonction génératrice des moments ϕ de la loi $\gamma_{p,\lambda}$. Calculer $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ et $\mathsf{Var}(X)$.
3. Montrer que le carré d'une variable aléatoire distribuée selon $\mathcal{N}(0,1)$ suit une loi gamma.
4. Calculer la fonction caractéristique χ de la loi $\gamma_{p,\lambda}$.
5. Soient $X_1, \ldots, X_k$ sont k variables aléatoires indépendantes, distribuées respectivement selon les lois $\gamma_{p_1,\lambda}, \ldots, \gamma_{p_k,\lambda}$. Déterminer la loi de $Y = \sum_{i=1}^k X_i$.
6. Soient $Z_1, \ldots, Z_d$ des variables aléatoires indépendantes distribuées selon la loi $\mathcal{N}(0,1)$. On note $Y = \sum_{i=1}^d Z_i^2$. Montrer que Y suit une loi gamma dont il faudra déterminer les paramètres. Cette loi s'appelle loi du χ^2.
7. Déterminer $x_0(d) := \arg\max \gamma_{d/2,1/2}(x)$. Trouver un équivalent de $\gamma_{d/2,1/2}(x_0(d))$ lorsque $d \to \infty$.
8. Pour Y comme à la question 6, la limite en loi de $\frac{Y-d}{\sqrt{d}}$ lorsque $d \to \infty$ existe-t-elle ? Si oui, identifier la limite.

2 Rappels sur les différents types de convergence

Exercice 4
1. Donner un exemple de suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires réelles qui converge vers 0 en probabilité mais ne converge pas presque sûrement vers 0.
2. Donner un exemple de suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires réelles qui converge vers 0 dans L^1 mais pas dans L^2.
3. Montrer que de toute suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires qui converge vers 0 en probabilité, on peut extraire une sous-suite $(X_{n_k})_{k\in\mathbb{N}}$ qui converge presque sûrement vers 0 lorsque $k \to \infty$.
4. Donner un exemple de suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires réelles qui converge presque sûrement vers 0 mais pas dans L^1.
5. Donner un exemple de suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires réelles qui converge en loi mais pas en probabilité.

Exercice 5 Soient $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ deux suites de variables aléatoires.
1. Montrer que si $X_n \to X$ et $Y_n \to Y$ en probabilité, alors $X_n + Y_n \to X + Y$, $aX_n + bY_n \to aX + bY$ (où a et b sont des constantes réelles), $X_nY_n \to XY$ en probabilité.
2. Montrer que si $X_n \to c$ (où c une constante) en loi, alors $X_n \to c$ en probabilité.
3. Donner un exemple de deux suites $(X_n)_{n\in\mathbb{N}}$ et $(Y_n)_{n\in\mathbb{N}}$ de variables aléatoires réelles qui convergent $X_n \to X$ et $Y_n \to Y$ en loi telles que $X_n + Y_n$ ne converge pas vers $X + Y$.

Exercice 6
1. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires distribuées selon $\mathcal{N}(m_n, \sigma_n^2)$. Quelles conditions doivent vérifier les suites numériques (m_n) et (σ_n) pour que la suite (X_n) soit tendue ?
2. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires distribuées selon $\mathcal{N}(m_n, \sigma_n^2)$. On suppose que $m_n \to m$ et $\sigma_n \to \sigma$. Montrer que X_n converge en loi vers $\mathcal{N}(m, \sigma^2)$.
3. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires distribuées selon $\mathcal{N}(0, \sigma_n^2)$. On suppose que $\sigma_n \to 0$. Déterminer la loi limite de X_n.
4. Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes de même loi μ. Montrer que $\frac{1}{n}\sum_{k=1}^n \delta_{X_k}$ converge étroitement presque sûrement vers μ.

3 Principes de l'estimation

Exercice 7 Lors d'un jeu télévisé, le présentateur utilise une machine qui génère des nombres aléatoires uniformément distribués sur l'intervalle $[0,\theta]$. Il se sert de cette machine $n = 10$ fois et présente la suite ainsi obtenue à 2 participants au jeu. Les joueurs doivent deviner θ ; celui qui s'approche le plus de θ gagne.

1. Décrire précisément l'espace statistique pour ce jeu.
2. Le joueur A utilise la statistique $S_n = \frac{2}{n}\sum_{k=1}^{n} X_k$. Est-ce un estimateur biaisé, cohérent ?
3. Le joueur B utilise la statistique $T_n = \max(X_k, k = 1,\dots,n)$. Est-ce un estimateur biaisé, cohérent ?
4. Si vous participez un jour à ce jeu, allez-vous utiliser S_n, T_n ou une meilleure statistique ?

4 Autour de la loi gaussienne

Exercice 8 Soient $X_1,\dots,X_n$ des variables aléatoires indépendantes et identiquement distribuées selon la loi $\mathcal{N}(m,\sigma^2)$, où m est connue tandis que σ est inconnue.

1. Proposer un estimateur « raisonnable » de σ^2.
2. Déterminer un intervalle de confiance de niveau α pour l'estimateur précédent.

Exercice 9

1. Soit $g_{0,V}$ la densité de la loi normale d-dimensionnelle $\mathcal{N}_d(0,V)$ de matrice de covariance V. Montrer que les ensembles de niveau h, pour $0 < h < ((2\pi)^d \det V)^{-\frac{1}{2}}$ sont des ellipsoïdes et déterminer les axes principaux.
2. Que faut-il changer pour la loi $\mathcal{N}_d(m,V)$?

Exercice 10 Soient μ_1 et μ_2 deux probabilités sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. On note λ_d la mesure de Lebesgue d-dimensionnelle et on suppose que les probabilités μ_1 et μ_2 admettent des densités f_1 et f_2 par rapport à la mesure λ_d. On définit

$$H(\mu_1;\mu_2) := \mu_1(\log\frac{f_1}{f_2}) = \begin{cases} \int_{\mathbb{R}^d} f_1(x)\log\frac{f_1(x)}{f_2(x)}\lambda_d(dx) & \text{si } \mu_1(f_2=0)=0, \\ +\infty & \text{sinon.} \end{cases}$$

Cette quantité, selon le contexte, s'appelle **entropie relative** ou **information de Kullback-Leibler**.

1. Soit $g : \mathbb{R}^d \to \mathbb{R}$, définie par la formule

$$g(x) := \begin{cases} \frac{f_1(x)}{f_2(x)} & \text{si } f_2(x) > 0, \\ 1 & \text{sinon.} \end{cases}$$

   Pour montrer que $H(\mu_1;\mu_2)$ est bien définie lorsque $\mu_1(f_2 = 0) = 0$, montrer qu'il suffit de considérer alors sans perte de généralité le cas où $f_1 = g f_2$.
2. Montrer que $H(\mu_1;\mu_2)$ existe dans $[0,+\infty]$. *(Indication : La fonction définie par la formule $\psi(s) = 1 - s + s\log s$, avec la convention $0\log 0 = 0$, est convexe sur $\mathbb{R}^*$.)*
3. Montrer que $H(\mu_1;\mu_2) = 0$ si et seulement si $\mu_1 = \mu_2$.
4. On considère la classe Π_C de mesures de probabilité sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ vérifiant
   – $\forall \mathbb{P} \in \Pi_C$ la mesure $\mathbb{P}$ est centrée et a comme matrice de covariance C.

– $\forall \mathbb{P} \in \Pi_C$ la mesure $\mathbb{P}$ admet une densité f par rapport à la mesure λ_d telle que

$$H(\mathbb{P}) := -\int_{\mathbb{R}^d} f(x)(\log f(x))\lambda_d(dx)$$

existe dans $\mathbb{R}$. Cette quantité porte le nom d'**entropie différentielle**.
Montrer qu'alors $H(\mathcal{N}_d(0,C)) = \sup\{H(\mathbb{P}), \mathbb{P} \in \Pi_C\}$.

5 Mesures radiales

Exercice 11 Soit $\mathbb{P}$ une probabilité sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$. La probabilité est dite **radiale** si elle reste invariante sous toute transformation orthogonale T, i.e. $\mathbb{P} \circ T^{-1} = \mathbb{P}$.

1. Proposer une définition plausible pour la probabilité uniforme sur la sphère unité d-dimensionnelle.
2. Vérifer que la probabilité ainsi définie est radiale.

6 Estimation non-paramétrique

Exercice 12 (Ordre stochastique) Soient $\mathbb{P}_1$ et $\mathbb{P}_2$ deux lois sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ dont les fonctions de répartition sont notées respectivement F_1 et F_2. On dit que $\mathbb{P}_1$ **minore stochastiquement** $\mathbb{P}_2$, et on note $\mathbb{P}_1 \preceq \mathbb{P}_2$ si pour tout $x \in \mathbb{R}$, on a $\mathbb{P}_1([x,\infty[) \leq \mathbb{P}_2([x,\infty[)$. Montrer que les affirmations suivantes sont équivalentes :

1. $\mathbb{P}_1 \preceq \mathbb{P}_2$.
2. Il existe deux variables aléatoires X_1 et X_2, définies sur un espace de probabilité approprié $(\Omega, \mathscr{F}, \mathbb{P})$, telles que $\mathbb{P} \circ X_1^{-1} = \mathbb{P}_1$, $\mathbb{P} \circ X_2^{-1} = \mathbb{P}_2$ et $\mathbb{P}(X_1 \leq X_2) = 1$.
3. Pour toute fonction $f : \mathbb{R} \to \mathbb{R}$ croissante bornée, $\mathbb{E}(f(X_1)) \leq \mathbb{E}(f(X_2))$.

Exercice 13 (Distance entre répartitions) Soient $p \geq 1$ réel et $\mathscr{F}_p$ l'ensemble de fonctions de répartition ayant de moments d'ordre p finis. Pour $F, G \in \mathscr{F}_p$ on introduit :

$$d_p(F,G) = \inf(\mathbb{E}(|X-Y|^p))^{1/p},$$

où l'infimum est calculé sur les paires des variables aléatoires (X,Y) dont la loi conjointe admet comme marginales respectives F et G.

1. Montrer que d_p est une distance.
2. Montrer que $d_1(F,G) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt$.
3. Montrer que $d_1(F,G) = \int_{\mathbb{R}} |F(x) - G(x)|dx$.
4. Soient $(G_n)_n$ une suite de $\mathscr{F}_p$ et $G \in \mathscr{F}_p$, pour $p > 1$. Quelle est la signification de $\lim_{n\to\infty} d_p(G_n, G) = 0$?

Exercice 14 (Estimation de la densité) Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées sur $\mathbb{R}$, dont la loi admet une densité f par rapport à la mesure de Lebesgue. On note F_n la fonction de répartition empirique et $f_n(t) = \frac{F_n(t+\lambda_n) - F_n(t-\lambda_n)}{2\lambda_n}$, pour $t \in \mathbb{R}$, où $(\lambda_n)_n$ est une suite de constantes numériques strictement positives.

1. Montrer que f_n est une densité.
2. Supposer que f soit continûment différentiable, $\lim \lambda_n = 0$ et $\lim(n\lambda_n) = \infty$. Montrer que $f_n(t)$ est un estimateur de $f(t)$ dont l'erreur quadratique est $\frac{f(t)}{2n\lambda_n} + o(\frac{1}{n\lambda_n}) + O(\lambda_n^2)$.
3. Si $\lim(n\lambda_n^3) = 0$ et les conditions de la question précédentes restent valables, montrer que

$$\sqrt{n}\lambda_n |f_n(t) - f(t)| \xrightarrow{\text{loi}} \mathcal{N}(0, \frac{f(t)}{2}).$$

4. Supposer f continue sur $[a,b]$ avec $-\infty < a < b < \infty$, $\lim \lambda_n = 0$ et $\lim(n\lambda_n^2) = \infty$. Montrer que $\int_a^b f_n(t) \to \int_a^b f(t)dt$ en probabilité.

Exercice 15 (Estimation avec des données manquantes) Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées sur $\mathbb{R}$, dont la loi admet comme fonction de répartition F. Soit $\pi(x) = \mathbb{P}(\delta_i = 1 | X_i = x)$ où

$$\delta_i = \begin{cases} 1 & \text{si } X_i \text{ est observé} \\ 0 & \text{si } X_i \text{ manque.} \end{cases}$$

Supposer que $0 < \rho = \int \pi(x) dF(x) < 1$.

1. Noter $G(x) = \mathbb{P}(X_i \leq x | \delta_i)$. Montrer que $F = G$ si et seulement si $\pi(x) = \rho$ pour tout x.
2. Soit $\hat{F}$ la fonction de répartition empirique qui place une masse $1/r$ à chaque X_i observé, où r est le nombre total des (X_i) observés. Montrer que $\hat{F}$ est un estimateur non-biaisé et cohérent de G.
3. Si π est constante, montrer que $\hat{F}$ est un estimateur non-biaisé et cohérent de F. Si π n'est pas constante, montrer que $\hat{F}$ est un estimateur biaisé et incohérent de F.

Exercice 16 (Densité du quantile empirique) Soit F_n la distribution empirique d'un échantillon de n variables aléatoires indépendantes et identiquement distribuées, dont la loi admet comme fonction de répartition F et comme densité f. On note $\phi_n(t)$ la densité du quantile empirique d'ordre p. Montrer que

$$\phi_n(t) = n C_{n-1}^{l_p - 1} (F(t))^{l_p - 1} (1 - F(t))^{n - l_p} f(t),$$

où

$$l_p = \begin{cases} np & \text{si } np \text{ est un entier} \\ 1 + \lfloor np \rfloor & \text{si } np \text{ n'est pas un entier.} \end{cases}$$

Exercice 17 (Statistique de Wilcoxon) Soient un échantillon de $n+m$ variables aléatoires indépendantes $(X_i)_{i=1,\ldots,n+m}$ à valeurs dans l'espace $(\mathbb{X}, \mathscr{X})$, où $\mathbb{X}$ est supposé totalement ordonné. On suppose que les n premières variables $(X_i)_{i=1,\ldots,n}$ sont distribuées selon une loi de fonction de répartition F et les m suivantes $(X_i)_{i=n+1,\ldots,n+m}$ selon une loi de fonction de répartition G. Les fonctions F et G sont supposées continues en tout point. On note $\pi \in S_{n+m}$ la permutation ordonnatrice pour l'échantillon complet et $\rho_i = \pi^{-1}(i)$ le rang de X_i.

1. Calculer $\sum_{i=1}^n \rho_i + \sum_{i=n+1}^{n+m} \rho_i$ et conclure qu'il suffit de considérer la statistique $W_F = \sum_{i=1}^n \rho_i$, la statistique $W_G = \sum_{i=n+1}^{n+m} \rho_i$ se déduisant trivialement de la précédente.
2. Montrer que $W_F = \frac{n(n+1)}{2} + U$ où $U := U_{n,m} = \sum_{i=1}^n \sum_{j=n+1}^{n+m} \mathbb{1}_{X_i > X_j}$.

3. On note pour C entier

$$A(C,n,m) := \mathsf{card}\{(c_1,\ldots,c_n) \in \mathsf{Ord}(\{0,\ldots,m\}^n) : \sum_{i=1}^{n} c_i = C\}.$$

Montrer que
– si $C<0$ alors $A(C,n,m)=0$;
– pour $C \geq 0$, on a $A(C,n,m) = A(C,m,n)$;
– $A(nm-C,n,m) = A(C,n,m)$ et
– pour $C \geq 0$, on a $A(C,n,m) = \sum_{j=0}^{n} A(C-j,j,m-1)$.

4. Si on note $\mathbb{P}^{\otimes(n+m)}$ la loi produit dans le cas $F=G$, montrer qu'alors

$$\mathbb{P}^{\otimes(n+m)}(U=c) = \frac{A(c,n,m)}{C_n^{n+m}}.$$

Cette expression permet de calculer explicitement la loi de U pour différentes valeurs de c,n,m, par récurrence. (Ce calcul n'est pas demandé.) Notons simplement que si l'on définit, pour $\alpha \in]0,1[$,

$$u_{n,\alpha} = \max\{c \in \mathbb{N} : \mathbb{P}^{\otimes 2n}(U \leq c) \leq \alpha\},$$

alors on obtient le tableau des quantiles ci-dessous.

n	4	5	6	7	8	9	10	11	12
$u_{n,0.05}$	2	5	8	12	16	21	28	35	42

Exercice 18 (Le cholesterol varie-t-il avec l'âge ?) Dans une étude clinique, on a mesuré le taux de cholesterol de 22 hommes dont 11 sont dans la classe d'âge 20–30 ans et 11 dans la classe 40–50. Le tableau suivant reproduit les valeurs mesurées. (Des lignes blanches sont laissées dans ce tableau pour votre convenance.)

20–30	135	222	251	260	269	235	386	252	352	173	156
rang											
40–50	294	311	286	264	277	336	208	346	239	172	254
rang											

1. Calculer le rang de chaque témoin dans l'échantillon global.
2. Calculer la statistique U de l'échantillon (voir exercice précédent).
3. Pour tester l'hypothèse $H_0 : F = G$ contre son alternative $H_1 : F > G$ proposer comme région de rejet $\phi = 1$ si $U < c$. Quelle est la valeur de c pour un niveau de signification de 5% ? (voir exercice précédent).
4. Quelle est votre conclusion au vu des résultats cliniques ?

7 Fondements de la statistique

7.1 Familles exponentielles

Exercice 19 (Double exponentielle) Soit Π la famille de probabilités sur $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ ayant une densité $f_\theta(x) = \frac{1}{2}\exp(-|x-\theta|)$, $\theta \in \mathbb{R}$, par rapport à la mesure de Lebesgue. Montrer que la famille n'est pas exponentielle.

Exercice 20 (Familles « position-échelle ») Définition : Soit $\mathbb{P}$ une probabilité sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Pour $V \subseteq \mathbb{R}^k$ et $\mathscr{M}_k$ une classe de matrices $k \times k$ symétriques et définies positives, on considère la famille de probabilités $\{P_{(m,\Sigma)}, m \in V, \Sigma \in \mathscr{M}_k\}$ où $P_{(m,\Sigma)}(B) = \mathbb{P}(\Sigma^{-1/2}(B-m))$ pour $B \in \mathscr{B}(\mathbb{R}^k)$. Cette famille est dite famille de **position-échelle**, où m est le paramètre de position et Σ le paramètre d'échelle.

Soit X une variable aléatoire distribuée selon la loi $\gamma_{p,\lambda}$. Il est rappelé que cette loi a une densité par rapport à la mesure de Lebesgue

$$\gamma_{p,\lambda}(x) = \frac{\lambda}{\Gamma(p)} x^{p-1} \exp(-\lambda x) \mathbb{1}_{[0,\infty[}(x).$$

On note $Y = \sigma \log X$. Montrer que si

1. le paramètre $\sigma > 0$ est inconnu, la loi de Y appartient à une famille position-échelle,
2. le paramètre $\sigma > 0$ est connu, la loi de Y appartient à une famille exponentielle.

Exercice 21 Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées selon la loi uniforme sur $[0,1]$. On note $S_n := X_{n:n} - X_{1:n}$.

1. Déterminer la densité de la loi de S_n.
2. Montrer que $2n(1-S_n) \xrightarrow{\text{loi}} \chi^2_4$. *(Indication : Pour cette deuxième question, commencer par établir le théorème de Scheffé : Si (f_n) est une suite de densités par rapport à une mesure ν sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ qui converge ν-p.p. vers une fonction f qui est aussi une densité par rappport à ν, alors $\int |f_n(x) - f(x)| \nu(dx) \to 0$.)*

Exercice 22 Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées selon la loi exponentielle de densité $f_X(x) = \frac{1}{\theta} \exp(-(a-x)/\theta) \mathbb{1}_{[0,\infty[}(x)$. On définit $X_{0:n} := 0$ et $Z_i := X_{i:n} - X_{i-1:n}$ pour $i = 1, \ldots, n$. Montrer que

1. $Z_1, \ldots, Z_n$ sont indépendantes et $(2n-i+1)Z_i/\theta)$ est distribuée selon χ^2_2.
2. $2[\sum_{i=1}^r X_{i:n} + (n-r)X_{r:n} - na]/\theta$ est distribuée selon χ^2_{2r}, pour $r = 1, \ldots, n$.
3. Si $Y = \frac{1}{n-1} \sum_{i=1}^n (X_{i:n} - X_{1:n})$, montrer que $X_{1:n}$ et Y sont indépendantes et $\frac{X_{1:n}-a}{Y}$ a une densité $n(1 + \frac{nx}{n-1})^{-n} \mathbb{1}_{[0,\infty[}(x)$.

(Indication : Il suffit d'établir les résultats pour $a = 0$ et $\theta = 1$.)

7.2 Théorie de la décision : admissibilité, optimalité

Exercice 23 (Calcul du risque) Soit $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ un echantillon de variables aléatoires indépendantes et identiquement distribuées selon la loi exponentielle de moyenne $\theta \in]0, \infty[$. Soit $\theta_0 > 0$ un paramètre fixé. On veut tester l'hypothèse nulle $H_0 : \theta \leq \theta_0$ contre son alternative $H_1 : \theta > \theta_0$ à l'aide de la règle de décision déterministe $D(X) = \mathbb{1}_{[c,\infty[}(\overline{X})$ où $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$ et la fonction de perte $l_\theta(\delta) = (1-\delta)\mathbb{1}_{\theta > \theta_0} + \delta \mathbb{1}_{\theta \leq \theta_0}$ pour $\delta \in \mathbb{D} = \{0,1\}$. Calculer le risque associé.

Exercice 24 (Les règles admissibles ne sont pas nécessairement bonnes) Soit $(\mathbb{X}, \mathscr{X}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ un modèle paramétrique. Supposons que pour un $\theta_0 \in \Theta$, le modèle est dominé par la probabilité $\mathbb{P}_{\theta_0}$. Montrer que la règle de décision $D \equiv \theta_0$ est admissible pour $l_\theta(\delta) = (\theta - \delta)^2$.

Exercice 25 Soit $(X_1,\dots,X_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(X_1^2)<\infty$. Noter $m(\mathbb{P})=\mathbb{E}X_1$ et considérer la fonction de perte quadratique $l_{\mathbb{P}}(\delta)=(m(\mathbb{P})-\delta)^2$. On introduit la règle $D(X)=\frac{1}{n}\sum_{i=1}^n X_i$ et pour a et b constantes la famille des règles $D_{a,b}(\cdot)=aD(\cdot)+b$.

1. Pour $a>1$, la règle $D_{a,b}$ est-elle admissible pour l'estimation de m ?
2. Pour $b\neq 0$, la règle $D_{1,b}$ est-elle admissible pour l'estimation de m ?

Exercice 26 Considérer le problème d'estimation d'un paramètre $\theta\in[a,b]\subseteq\mathbb{R}$, où a et b connus. Supposons que $[a,b]\subseteq\mathbb{D}$ et que la fonction de perte s'écrive comme $l_\theta(\delta)=L(|\theta-\delta|)$, où $L:[0,\infty[\to\mathbb{R}_+$ est strictement croissante. Montrer que toute règle de décision D telle que $\mathbb{P}(D(X)\notin[a,b])>0$, pour un $\mathbb{P}\in\Pi$ est inadmissible.

Exercice 27 Soit $(X_1,\dots,X_n)$ un échantillon de n variables aléatoires ayant toutes la même espérance et sont toutes de carré intégrables, i.e. pour tout $i=1,\dots,n$ on a $m(\mathbb{P}):=\mathbb{E}(X_i)$ et $\sigma_i^2(\mathbb{P})=\mathbb{E}(X_i^2)<\infty$. Considérer le problème d'estimation de $m(\mathbb{P})$ comme un problème de décision associé à une fonction de perte quadratique $l_{\mathbb{P}}(\delta)=(m(\mathbb{P})-\delta)^2$.

1. Montrer que si $\mathscr{C}$ contient toutes les règles de décision, il n'existe pas de règle $\mathscr{C}$-opitmale.
2. Si $\sigma_i^2=\sigma^2/a_i$, avec a_i connu pour $i=1,\dots,n$ et σ^2 inconnu, trouver une règle $\mathscr{C}$-optimale pour

$$\mathscr{C}=\{D:\mathbb{X}^n\to\mathbb{R}:D(x)=\sum_{i=1}^n c_ix_i;c_i\in\mathbb{R}_+;\sum_{i=1}^n c_i=1\},$$

   pour des variables $(X_1,\dots,X_n)$ non-corrélées.
3. Pour $\mathscr{C}$ comme à la question précédente et $(X_1,\dots,X_n)$ identiquement distribuées avec coefficient de corrélation ρ, déterminer une règle $\mathscr{C}$-optimale.

7.3 Exhaustivité

Définition (Exhaustivité minimale) : Soient $(\mathbb{X},\mathscr{X},(\mathbb{P}_\theta)_{\theta\in\Theta})$ un espace statistique et $S:(\mathbb{X},\mathscr{X})\to(\mathbb{D},\mathscr{D})$ une statistique. La statistique S est dite **exhaustive minimale** si elle est exhaustive et pour toute statistique S' exhaustive, il existe une application mesurable ψ telle que $S=\psi\circ S'$.

Exercice 28 (Un critère d'exhaustivité minimale) Soient $(\mathbb{X},\mathscr{X},(\mathbb{P}_\theta)_{\theta\in\Theta})$ un espace statistique et $S:(\mathbb{X},\mathscr{X})\to(\mathbb{D},\mathscr{D})$ une statistique. On suppose le modèle dominé par une probabilité μ et on note $f_\theta=\frac{d\mathbb{P}_\theta}{d\mu}$. Montrer que si on a l'équivalence

$$S(x)=S(y)\Leftrightarrow\theta\mapsto\frac{f_\theta(x)}{f_\theta(y)}\text{ indépendante de }\theta,$$

alors S est exhaustive minimale pour θ.

Exercice 29 Soit un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon la loi $\mathrm{unif}(]\theta,\theta+1[)$, pour un $\theta\in\mathbb{R}$. Montrer que la statistique $S:\mathbb{R}^n\to\mathbb{R}^2$ définie par $S(x_1,\dots,x_n)=(x_{1:n},x_{n:n})$ est minimalement exhaustive pour θ.

Exercice 30 Soient un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon une loi ayant comme densité par rapport à la mesure de Lebesgue

$$f_\theta(x) = \frac{1}{\theta}\exp(-\frac{x-\theta}{\theta})\mathbb{1}_{]\theta,\infty[}(x)$$

pour un $\theta > 0$. Trouver une statistique minimalement exhaustive pour θ.

Exercice 31 Soient X, Y deux variables aléatoires telles que Y soit distribuée selon $\mathscr{B}(N,\pi)$ et X admette comme loi conditionnelle sachant que $Y = y$ la loi $\mathscr{B}(y,p)$, avec $\pi, p \in]0,1[$.

1. Supposons que π, p soient des paramètres inconnus tandis que N est connu. Montrer que (X,Y) est minimalement exhaustive pour (π,p).
2. Supposons que π, N soient des paramètres connus tandis que p est inconnu.
   – La statistique X est-elle minimalement exhaustive pour p ?
   – La statistique Y est-elle minimalement exhaustive pour p ?

Exercice 32 Soit $\Pi = (\mathbb{P}_\theta)_{\theta\in\Theta}$ une famille de probabilités telles que, pour tout θ, la probabilité $\mathbb{P}_\theta$ admette une densité $f_\theta > 0$ par rapport à la mesure de Lebesgue. On suppose en outre que pour tout θ la densité $f_\theta(x)$ est continue en x. Soient X_1 et X_2 deux variables aléatoires indepéndantes et identiquement distribuées selon f_θ. Montrer que si $X_1 + X_2$ est une statistique exhaustive, alors Π est une famille exponentielle.

7.4 Complétude

Définition (Liberté) : Soient $(\mathbb{X}, \mathscr{X}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ un espace statistique et $S : (\mathbb{X}, \mathscr{X}) \to (\mathbb{D}, \mathscr{D})$ une statistique. Si la loi de S ne dépend pas de θ alors la statistique s'appelle **libre**.

Exercice 33 Soient un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon la loi unif$(]\theta,\theta+1[)$, pour un $\theta \in \mathbb{R}$. Soit $S : \mathbb{R}^n \to \mathbb{R}^2$ la statistique définie par $S(x_1,\dots,x_n) := (R(x), M(x)) \equiv (x_{n:n} - x_{1:n}, \frac{x_{1:n}+x_{n:n}}{2})$. Montrer que S est minimalement exhaustive pour θ tandis que R est libre.

Exercice 34 (Famille paramétrique de position) Soit un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon une loi ayant comme fonction de répartition $F(x-\theta)$, $\theta \in \mathbb{R}$. Montrer que $R = X_{n:n} - X_{1:n}$ est libre.

Exercice 35 (Famille paramétrique d'échelle) Soit un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon une loi ayant comme fonction de répartition $F(x/\theta)$, $\theta > 0$. Monter que toute statistique qui dépend de l'échantillon à travers les $n-1$ variables aléatoires $X_1/X_n, \dots, X_{n-1}/X_n$ est libre.

Exercice 36 Soient X_1, X_2 deux variables aléatoires indépendantes distribuées selon la loi discrète :

$$\mathbb{P}_\theta(X_1 = \theta) = \mathbb{P}_\theta(X_1 = \theta+1) = \mathbb{P}_\theta(X_1 = \theta+2) = 1/3, \text{ avec } \theta \in \mathbb{Z}.$$

La statistique (R,M) est minimalement exhaustive tandis que R est libre. Argumenter comment R, en conjonction avec M, peut fournir de l'information sur θ.

Exercice 37 (Statistique binomiale exhaustive complète) Suppposons que S ait une loi binomiale $\mathscr{B}(n,\theta)$, avec $\theta \in]0,1[$. Soit g une fonction telle que $\mathbb{E}_\theta g(S) = 0$, pour tout θ. Monter que $\mathbb{P}_\theta(g(S)=0) = 1$ pour tout $\theta \in]0,1[$.

Exercice 38 (Statistique uniforme exhaustive complète) Soit $(X_1,\dots,X_n)$ un échantillon de n variables aléatoires indépendantes identiquement distribuées selon $\mathsf{unif}(]0,\theta[)$, avec $0 < \theta < \infty$. Soit S la statistique $S(x) := x_{n:n}$. Montrer que
– la statistique S est exhaustive,
– la statistique S est complète.

Exercice 39 (Théorème de Basu) Montrer que si une statistique S est complète et minimalement exhaustive alors S est indépendante de toute statistique libre. (Donner la démonstration uniquement dans le cas discret).

Exercice 40 (Application du théorème de Basu) Soit $(X_1,\dots,X_n)$ un échantillon de n variables aléatoires indépendantes identiquement distribuées selon $\mathsf{expon}(\theta)$, avec $0 < \theta < \infty$.
– Monter que la famille des lois exponentielles forme une famille d'échelle.
– En conclure que $g(X) = \frac{X_n}{X_1+\dots+X_n}$ est une statistique libre.
– Montrer que la famille des lois exponentielles forme une famille exponentielle et déterminer sa forme.
– En conclure que la statistique $S(X) = \sum_{i_1}^n X_i$ est une statistique exhaustive et complète.
– Montrer que $g(X)$ et $S(X)$ sont indépendantes.

8 Théorie des estimateurs ponctuels

Exercice 41 Montrer que pour un modèle statistique dominé, l'entropie de Kullback-Leibler vérifie

$$H(\mathbb{P}_{\theta_1},\mathbb{P}_{\theta_2}) = \frac{1}{2}\langle \theta_2-\theta_1, I(\theta_1)(\theta_2-\theta_1)\rangle + o(\|\theta_1-\theta_2\|^2),$$

où $I(\theta_1)$ est la matrice d'information deFisher.

Exercice 42 (Des estimateurs efficaces n'existent pas toujours) Soit $(X_1,\dots,X_n)$ un échantillon de variables aléatoires indépendantes et identiquement distribuées sur $(\mathbb{X},\mathscr{X},\Pi)$ où $\Pi = \{\mathscr{N}(m,\sigma^2), m \in \mathbb{R}, \sigma^2 > 0\}$. Considérer la statistique S^2 variance empirique et la statistique cS^2 avec $c > 0$. Argumenter.

Exercice 43 Soit un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon une loi normale $\mathscr{N}(\theta,1)$. Écrire l'équation de maximum de vraisemblance et déterminer une solution.

Exercice 44 Soit un échantillon de n variables aléatoires indépendantes $(X_i)_{i=1,\dots,n}$ identiquement distribuées selon une loi binomiale $\mathscr{B}(\theta),p)$ de taille $\theta \in \mathbb{N}^*$ inconnue et paramètre p connue.
– Écrire l'équation du maximum de vraisemblance.
– Pour $p = 1/2$ et $n = 4$, déterminer le maximum pour l'observation $X_1 = 0, X_2 = 20, X_3 = 1, X_4 = 19$.

9 Bibliographie

La référence de base est indiquée en **caractères gras**.

P. Billingsley, Probability and measure. Third edition. Wiley Series in Probability and Mathematical Statistics. John Wiley, New York (1995).

D. Dacunha-Castelle, M. Duflo, Probabilités et statistique, vol. 1, Masson, Paris (1982).

D. Fourdrinier, Statistique inférentielle, Dunod, Paris (2002).

G. Keller, Vorlesung mathematische Statistik, Vorlesungsskript Universität Erlangen/Nürnberg (1992).

E.L. Lehmann, Testing statistical hypotheses, Chapman and Hall, , New York (1986).