

1 Rappels sur la loi normale

Le théorème Central Limite nous informe sur le rôle essentiel de la loi normale en théorie statistique. Ce chapitre a pour but de (re)voir certaines propriétés et lois issues de la loi normale, à cause de leur importance en statistique.

Nous savons que si $X \sim N(\mu, \sigma^2)$, alors

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

($a \neq 0$) et que si $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ et X_1, X_2 sont indépendantes, alors

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

(ceci est-il vrai si X_1 et X_2 ne sont pas indépendantes?).

On généralise à plusieurs variables.

Théorème 1 Soit X_1, \dots, X_n indépendantes et telles que $X_i \sim N(\mu_i, \sigma_i^2)$. Alors la variable

$$Y = \sum_{i=1}^n a_i X_i + b \sim N\left(\mu = \sum_{i=1}^n a_i \mu_i + b, \sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

(on suppose $\sum_{i=1}^n a_i^2 \neq 0$).

Preuve. Utilisons les FGM. Nous savons que si $V \sim N(\mu, \sigma^2)$, alors

$$M_V(t) = \exp(\mu t + t^2 \sigma^2 / 2)$$

Maintenant,

$$\begin{aligned} M_Y(t) &= E\left(\exp\left(\sum_{i=1}^n a_i X_i + b\right)t\right) = E\left(e^{bt} \prod_{i=1}^n e^{a_i X_i t}\right) \stackrel{ind.}{=} e^{bt} \prod_{i=1}^n M_{X_i}(a_i t) \\ &= e^{bt} \prod_{i=1}^n \exp(a_i \mu_i t + a_i^2 t^2 \sigma_i^2 / 2) = \exp\left(t\left(b + \sum_{i=1}^n a_i \mu_i\right) + t^2 \sum_{i=1}^n a_i^2 \sigma_i^2 / 2\right) \end{aligned}$$

On reconnaît la FGM d'une variable $N(\mu, \sigma^2)$. ■

Corollaire 1 Si X_1, \dots, X_n sont iid $\sim N(\mu, \sigma^2)$, alors

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2 / n)$$

Rappelons aussi le résultat suivant.

Théorème 2 Soit X_1, \dots, X_n indépendantes et telles que $X_i \sim N(\mu_i, \sigma_i^2)$. Soit les variables normales

$$U = \sum_{i=1}^n a_i X_i$$

et

$$V = \sum_{i=1}^n b_i X_i$$

Alors

1.

$$Cov(U, V) = \sum_{i=1}^n a_i b_i \sigma_i^2$$

2. U et V sont indépendantes si et seulement si

$$Cov(U, V) = 0$$

1.1 Exercices

Exercice 1 Soit $X_1 \sim N(\mu = 3, \sigma^2 = 5)$ et $X_2 \sim N(\mu = -7, \sigma^2 = 2)$, indépendantes.

1. Quelle est la loi de $U = 4X_1 - X_2/3$?
2. Calculer $\text{Cov}(U, V)$ où $V = X_1$.

Exercice 2 Soit $X \sim N(0, 1)$ et Y telle que $P(Y = \pm 1) = 1/2$. On supposera X et Y indépendantes et posons $V = XY$.

1. Vérifier que $V \sim N(0, 1)$.
2. Vérifier que $\text{Cov}(X, V) = 0$.
3. Vérifier que X et V sont indépendantes.
4. Ce résultat contredit-il le théorème 2 ?

2 Loi du Khi-carré, loi de Student et loi F

Les lois présentées ici ne servent en général pas comme modèles de phénomènes qu'on veut décrire/simuler. Leur importance est due au fait qu'elles interviennent dans l'analyse des intervalles de confiance et des tests d'hypothèses.

2.1 Loi du Khi-carré

Soit $Z \sim N(0, 1)$. Alors la variable aléatoire $Y = Z^2$ suit une loi du Khi-carré à un degré de liberté : $Y \sim \chi_1^2$.

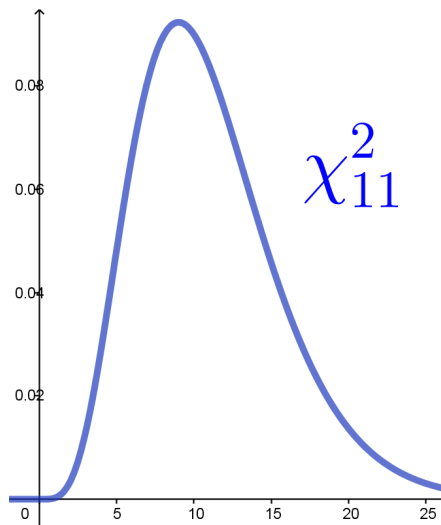
On peut faire les observations suivantes.

1. La fonction de densité de Y est

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}, \quad y \geq 0$$

En effet, la fonction de répartition de Y est

$$\begin{aligned} F_Y(y) &= P(Z^2 \leq y) = P(-\sqrt{y} \leq Z \leq \sqrt{y}) \quad (y \geq 0) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1 \\ \Rightarrow f_Y(y) &= \frac{2}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} = \text{résultat} \end{aligned}$$



2. C'est donc un cas particulier de la distribution gamma : $\chi_1^2 = \text{Gamma}(\alpha = \frac{1}{2}, \lambda = \frac{1}{2})$

3. Sa FGM est

$$M(t) = (1 - 2t)^{-1/2}$$

4. Si $X \sim N(\mu, \sigma^2)$, alors $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ et $Y = Z^2 \sim \chi_1^2$

5. Si Y_1, \dots, Y_n sont iid χ_1^2 alors $V = Y_1 + \dots + Y_n \sim \chi_n^2 = \text{Gamma}(\alpha = \frac{n}{2}, \lambda = \frac{1}{2})$. De plus,

$$\begin{aligned} E(V) &= n \\ \text{VAR}(V) &= 2n \\ M_V(t) &= (1 - 2t)^{-n/2} \\ f_V(v) &= \frac{v^{(n/2)-1}}{2^{n/2}\Gamma(n/2)} e^{-v/2}, \quad v \geq 0 \\ E(V^k) &= \frac{2^k \Gamma(\frac{n}{2} + k)}{\Gamma(\frac{n}{2})}, \quad k \geq 0 \end{aligned}$$

En particulier, si $n = 2$, $V \sim \text{Exp}(\lambda = 1/2)$.

6. Si $V \sim \chi_n^2$ et $W \sim \chi_m^2$ sont indépendantes, alors $V + W \sim \chi_{n+m}^2$.

7. Commandes R : `pchisq(x,n)` ($= P(\chi_n^2 \leq x)$), `dchisq(x,n)` ($= f(\chi_n^2 = x)$), `qchisq(x,n)`.

Exemple 1 Les commandes R suivantes permettent d'imprimer une table typique des quantiles du Khi carré pour les probabilités 0.01, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.99 pour les degrés de liberté $n = 1, \dots, 30$:

```
n <- 1 :30; p <- c(.01,.025,.05,.1,.9,.95,.975,.99);
for(r in n) {print(c(r, round(qchisq(p,r),digits = 3)))}
```

2.1.1 Exercices

Exercice 3 Avec les hypothèses ci-dessus (propriété 6), est-il vrai que $V - W \sim \chi_{n-m}^2$?

Rappel. Si $X \sim \text{Gamma}(\alpha, \lambda)$, alors

$$\begin{aligned} f_X(x) &= \frac{\lambda(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x}, \quad x \geq 0, \alpha, \lambda > 0 \\ M_X(t) &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha = (1 - t/\lambda)^{-\alpha} \end{aligned}$$

Exercice 4 Si $X \sim \text{Gamma}(n, \lambda)$, quelle est la loi de $Y = 2\lambda X$?

Exercice 5 Quelle serait, selon vous, une bonne définition d'une loi du Khi carré à zéro degré de liberté ?
Indication. Voir la FGM.

2.2 Loi de Student

Définition 1 Soit donné deux variables aléatoires indépendantes $Z \sim N(0, 1)$ et $U \sim \chi_n^2$. Alors la variable

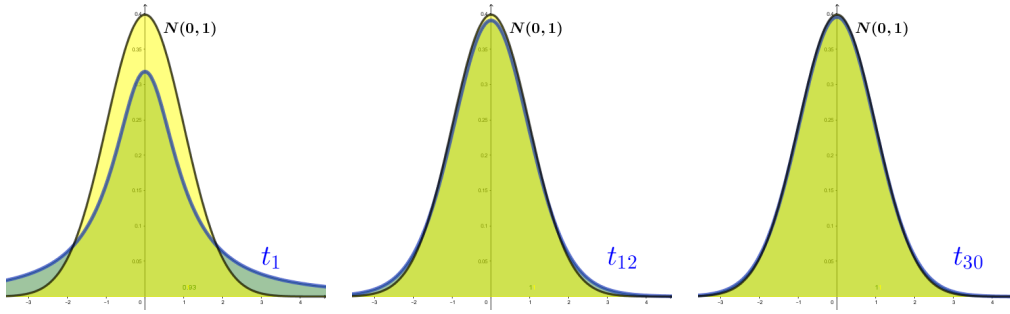
$$T = Z/\sqrt{U/n}$$

suit une loi de Student (ou loi t) à n degrés de liberté. On écrit

$$T \sim t_n$$

La densité de T est donnée par

$$f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad t \in \mathbb{R}$$



Propriétés

1. $f(-t) = f(t)$ (fonction paire : graphe symétrique par rapport à l'axe des y). Pour les petites valeurs de n , la 'queue' de la densité descend lentement vers l'axe des x (contrairement par exemple à celle de $N(0, 1)$ qui contient l'essentiel de la probabilité au bout de 3 unités).
2. $E(T) = 0$ ($n \geq 2$), $VAR(T) = \frac{n}{n-2}$ ($n > 2$).
3. Si $n \uparrow \infty$, la distribution se stabilise ($N(0, 1)$) (convergence en loi).
4. Si $n = 1$, $T \sim Cauchy$. Pas de moyenne ni de variance.
5. Commandes R : `dt(x,n)`, `pt(x,n)`, `qt(q,n)`.

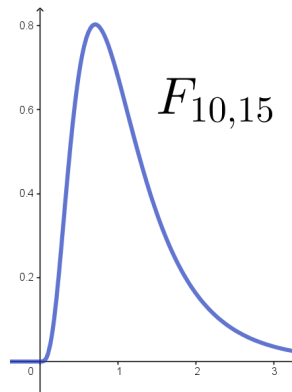
2.3 Loi F (Fisher-Snedecor)

Définition 2 Soit donné deux variables aléatoires indépendantes $U \sim \chi_m^2$ et $V \sim \chi_n^2$. Alors la variable

$$W = \frac{U/m}{V/n} \sim F_{m,n}$$

(loi de Fisher à m et n degrés de liberté). La densité de W est donnée par

$$f(w) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2}, \quad w > 0$$



Propriétés

1. $E(W) = \frac{n}{n-2}$ ($n > 2$), $VAR(W) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ ($n > 4$).
2. Si $W \sim F_{m,n}$ alors $1/W \sim F_{n,m}$.
3. Si $T \sim t_n$ alors $T^2 \sim F_{1,n}$.
4. Si $W \sim F_{m,n}$ alors $mW \xrightarrow{Loi} \chi_m^2$ (quand $n \rightarrow \infty$).
5. Soit $X_1, \dots, X_n \text{ iid} \sim N(\mu_X, \sigma^2)$ et $Y_1, \dots, Y_m \text{ iid} \sim N(\mu_Y, \sigma^2)$ telles que les X_i soient indépendants de Y_j , alors

$$F = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}{\frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2} \sim F_{n-1, m-1}$$

6. Commandes R : `df(x,m, n)`, `pf(x,m, n)`, `qf(q,m, n)`.

3 Lois de \bar{X} et S^2

Soit X_1, \dots, X_n iid $N(\mu, \sigma^2)$. On définit la moyenne et la variance empiriques (échantillonales)

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

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

Un petit calcul montre que

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

En effet,

$$\begin{aligned} (X_i - \bar{X})^2 &= X_i^2 - 2X_i\bar{X} + \bar{X}^2 \\ \Rightarrow \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \sum_{i=1}^n X_i^2 - \sum_{i=1}^n 2X_i\bar{X} + \sum_{i=1}^n \bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 \\ &\Rightarrow \text{résultat} \end{aligned}$$

On a les résultats importants suivants.

Théorème 3 Avec les hypothèses et notations ci-dessus,

1. $\bar{X} \sim N(\mu, \sigma^2/n)$
2. $E(S^2) = \sigma^2$
3. $VAR(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right)$ où $\mu_4 = E(X^4)$

Preuve.

1. Évident.
- 2.

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} E \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = \frac{1}{n-1} E \left(\sum_{i=1}^n X_i^2 \right) - \frac{n}{n-1} E(\bar{X}^2) \\ &= \frac{1}{n-1} \sum_{i=1}^n E(X_i^2) - \frac{n}{n-1} E(\bar{X}^2) = \frac{n}{n-1} E(X_1^2) - \frac{n}{n-1} (\mu^2 + \sigma^2/n) \\ &= \frac{n}{n-1} (\mu^2 + \sigma^2) - \frac{n}{n-1} (\mu^2 + \sigma^2/n) = \frac{n}{n-1} (\mu^2 + \sigma^2 - \mu^2 - \sigma^2/n) \\ &= \frac{n}{n-1} (\sigma^2 - \sigma^2/n) = \sigma^2 \end{aligned}$$

3. Omise.

■

Théorème 4 Avec les hypothèses ci-dessus, on a

1. \bar{X} et $X_i - \bar{X}$ sont indépendantes, $i = 1, \dots, n$ (résultat non intuitif).
2. \bar{X} et S^2 sont indépendantes.
3. $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.
4. $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

Preuve.

1. Utilise la MGF conjointe (cf Rice ou remarque plus bas pour une approche directe).
2. Conséquence de 1. Mais donnons une preuve dans le cas $n = 2$. On a

$$\begin{aligned} X_i - \mu &= (X_i - \bar{X}) + (\bar{X} - \mu) \Rightarrow \sigma^2 S^2 = \frac{1}{2-1} \sum_{i=1}^2 (X_i - \bar{X})^2 \\ &= \left(X_1 - \frac{X_1 + X_2}{2} \right)^2 + \left(X_2 - \frac{X_1 + X_2}{2} \right)^2 \\ &= \left(\frac{1}{2} (X_1 - X_2) \right)^2 + \left(\frac{1}{2} (X_2 - X_1) \right)^2 = \frac{1}{2} (X_1 - X_2)^2 \\ &\Rightarrow \frac{S^2}{(2-1)\sigma^2} = \frac{(X_1 - X_2)^2}{2\sigma^2} = \left(\frac{X_1 - X_2}{\sqrt{2}\sigma} \right)^2 \end{aligned}$$

Maintenant,

$$\frac{X_1 - X_2}{\sqrt{2}\sigma} \sim N(0, 1) \Rightarrow \frac{S^2}{(2-1)\sigma^2} = \left(\frac{X_1 - X_2}{\sqrt{2}\sigma} \right)^2 \sim \chi_1^2$$

Par ailleurs, il est facile de voir (exercice) que

$$Y_1 = \frac{X_1 + X_2}{\sigma} \text{ et } Y_2 = \frac{X_1 - X_2}{\sigma}$$

sont indépendantes. Puisque

$$\bar{X} = \frac{\sigma}{2} Y_1$$

ne dépend que de Y_1 et

$$S^2 = \frac{\sigma^2}{2} Y_2^2$$

ne dépend que de Y_2 , alors \bar{X} et S^2 sont indépendantes.

3. On utilise l'égalité

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$$

Alors (petit calcul)

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

Le reste suit par une utilisation des FGM (voir exercice 3 ci-dessus).

4. On a

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)}{\sqrt{S^2/\sigma^2 \sim \chi_{n-1}^2}} \stackrel{\text{ind+déf}}{\sim} t_{n-1}$$

■

Remarque 1 La propriété 4 peut servir à la construction d'un intervalle de confiance pour une moyenne de variable $X_1, \dots, X_n \text{ iid } \sim N(\mu, \sigma^2)$, μ et σ^2 inconnus. En effet, soit $\alpha \in]0; 1[$ donné et soit $t_{n-1, \alpha/2}$ le nombre positif défini par

$$\begin{aligned} 1 - \alpha &= P(-t_{n-1, \alpha/2} \leq T \leq t_{n-1, \alpha/2}) = P\left(-t_{n-1, \alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1, \alpha/2}\right) \\ &= P\left(\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}\right) \end{aligned}$$

Remarque 2 Intervalle de prédiction. Soit X_1, \dots, X_n, X_{n+1} iid $N(\mu, \sigma^2)$, μ, σ^2 inconnus. On observe X_1, \dots, X_n et on utilise cette information pour prédire X_{n+1} . Puisque μ est inconnu, on utilisera (prédicteur naturel de X_{n+1})

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Rappelons que $\bar{X}_n \sim N(\mu, \sigma^2/n)$ et que X_{n+1} est indépendant de \bar{X}_n . Par conséquent,

$$\begin{aligned} X_{n+1} - \bar{X}_n &\sim N(0, \sigma^2(1 + 1/n)) \\ \Rightarrow Z &= \frac{X_{n+1} - \bar{X}_n}{\sigma\sqrt{1 + 1/n}} \sim N(0, 1) \\ \Rightarrow T &= \frac{X_{n+1} - \bar{X}_n}{S_n\sqrt{1 + 1/n}} \sim t_{n-1} \\ \Rightarrow P(-t_{\alpha/2, n-1} < T < t_{\alpha/2, n-1}) &= 1 - \alpha \\ \Rightarrow \boxed{P(\bar{X}_n - t_{\alpha/2, n-1} S_n \sqrt{1 + 1/n} < X_{n+1} < \bar{X}_n + t_{\alpha/2, n-1} S_n \sqrt{1 + 1/n})} &= 1 - \alpha \\ &IP_{1-\alpha} \end{aligned}$$

Remarque 3 *La propriété 1 peut être démontrée directement en utilisant une certaine transformation orthogonale dite transformation de Helmert. Posons

$$\begin{aligned} Y_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \\ Y_2 &= \frac{1}{\sqrt{2}} (X_1 - X_2) \\ Y_3 &= \frac{1}{\sqrt{6}} (X_1 + X_2 - 2X_3) \\ &\vdots \\ Y_n &= \frac{1}{\sqrt{n(n-1)}} (X_1 + X_2 + \dots + X_{n-1} - (n-1)X_n) \end{aligned}$$

ou les X_i sont iid $N(\mu, \sigma^2)$. La matrice de cette transformation linéaire est donc

$$A = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \dots & \frac{n-1}{\sqrt{n(n-1)}} \end{bmatrix}$$

On a ainsi

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$

où

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

Puisque la matrice A est orthogonale (exercice !), son inverse est égal à sa transposée. On en déduit

$$\sum_{i=1}^n X_i^2 = \mathbf{X}^t \mathbf{X} = \mathbf{X}^t A^t A \mathbf{X} = \mathbf{Y}^t \mathbf{Y} = \sum_{i=1}^n Y_i^2$$

Géométriquement, cela signifie qu'une hypersphère (sphère dans \mathbb{R}^n) dans le repère des x_i reste une sphère dans le repère des y_i quand on lui imprime une rotation orthogonale. Le jacobien de cette transformation est la matrice $J = A^t$ avec $|J| = 1$. Maintenant, la densité conjointe des X_i est donnée par

$$f(x_1, \dots, x_n) = (\sqrt{2\pi}\sigma)^{-n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^2 - 2\mu x_i + \mu^2) \right]$$

On en déduit (formule de transformation) la densité conjointe des Y_i :

$$\begin{aligned} g(y_1, \dots, y_n) &= f(x_1(y_1, \dots, y_n), \dots, x_n(y_1, \dots, y_n)) |J| \\ &= (\sqrt{2\pi}\sigma)^{-n} \exp \left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n y_i^2 - 2\mu\sqrt{n}y_1 + n\mu^2 \right) \right] \\ &= (\sqrt{2\pi}\sigma)^{-n} \exp \left[-\frac{1}{2\sigma^2} \left(\sum_{i=2}^n y_i^2 + (y_1 - \sqrt{n}\mu)^2 \right) \right] \end{aligned}$$

On voit que les Y_i sont indépendantes. De plus, $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$ tandis que $Y_2, \dots, Y_n \sim N(0, \sigma^2)$. A partir de ceci, la propriété 1 est immédiate (exercice).