Statistique (MA101) Cours 3

ENSTA 1ère année

Christine Keribin

christine.keribin@math.u-psud.fr

Laboratoire de Mathématiques
Université Paris-Sud

2017-2018

Dans les cours précédents

- Echantillon, modèle statistique paramétrique,
- Estimateur
  - $\hookrightarrow$ Méthodes de construction : moments, max. de vraisemblance
  - $\hookrightarrow$ Propriétés : biais, variance, risque, consistance, loi asymptotique
- Vrai ou Faux ?
  - $\hookrightarrow$ Soit $\mu = \mathbb{E}(X_1)$. $T_n = \sum_i (X_i - \mu)^2/n$ est un estimateur sans biais de $\sigma^2 = \mathrm{Var}(X_1)$
  - $\hookrightarrow$ Un estimateur non biaisé est de risque minimum
  - $\hookrightarrow$ Un estimateur dont la variance tend vers 0 est consistant

Pour aller plus loin...

- Un constructeur automobile indique une consommation de $c_0 = 6.32\ell/100km$ pour les véhicules d'un type donné, dans des conditions précises de roulage, avec un écart-type de $\sigma_0 = 0.21\ell/100km$
- Un organisme indépendant prend 30 véhicules au hasard, et les soumet aux conditions de roulage nominales. Il observe $\bar{x} = 6.43\ell/100km > c_0$, $\widehat{\sigma} = 0.25\ell/100km$.
  - $\hookrightarrow$ Est-ce dû à la variabilité naturelle de l'expérience ?
  - $\hookrightarrow$ Où le constructeur a-t-il sous-estimé la consommation de ses véhicules ?
- Déterminer si le fait d'observer une moyenne plus grande que 6.43 est d'une probabilité forte ou pas sous les indications du constructeur.
  - $\hookrightarrow$ accéder à $\mathbb{P}(\bar{X} \geq 6.43)$,
  - $\hookrightarrow$ loi de $\bar{X}$, mais aussi de $\widehat{\sigma}^2$, ...

Sommaire

Définitions

Définition (Loi gaussienne sur $\mathbb{R}$)

La loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$ est la probabilité de densité par rapport à la mesure de Lebesgue

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Définition (Vecteur Gaussien)

Un vecteur aléatoire Y à valeurs dans $\mathbb{R}^n$ est gaussien si et seulement si toute combinaison linéaire de ses coordonnées est gaussienne, ie :

$$\text{Pour tout } U \in \mathbb{R}^n, \ \exists \mu \in \mathbb{R}^n, \sigma^2 \in \mathbb{R}^+ \text{ t.q. } U'Y \sim \mathcal{N}(\mu_U, \sigma_U^2)$$

Propriétés

Soit Y un vecteur aléatoire gaussien de dimension n.

- Sa loi est complètement déterminée par
  - $\hookrightarrow$ Espérance :
    $\mathbb{E}(Y) = (\mathbb{E}(Y_1), \ldots, \mathbb{E}(Y_n))' = \mu \in \mathbb{R}^n$
  - $\hookrightarrow$ Variance : $\mathrm{Var}(Y) = (\mathrm{cov}(Y_i, Y_j)) = \Sigma$, matrice $n \times n$.
- Si Σ est inversible, sa densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ est

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp -\frac{(y-\mu)'\Sigma^{-1}(y-\mu)}{2}$$

- Si A est une matrice $p \times n$ et $Y \sim \mathcal{N}_n(\mu, \Sigma)$,

$$AY \sim \mathcal{N}_p(A\mu, A\Sigma A')$$

Propriétés

- Si Y est un vecteur gaussien et si sa variance est diagonale par blocs, alors les blocs de coordonnées correspondants forment des vecteurs gaussiens indépendants.
- Un n-échantillon gaussien centré réduit est un vecteur gaussien de loi $\mathcal{N}_n(0, Id_n)$, c'est-à-dire un vecteur dont les n composantes sont des variables aléatoires indépendantes de loi gaussienne centrée réduite.
- Lorsqu'on fait un changement de base orthonormée, un vecteur gaussien reste un vecteur gaussien.

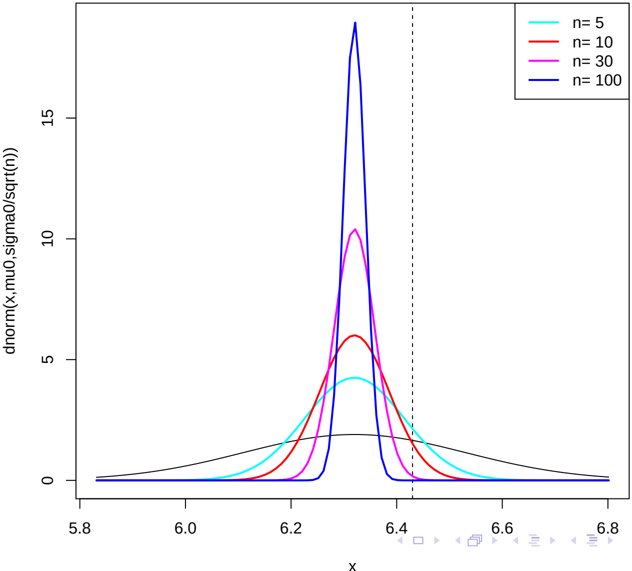
Loi de $\bar{X}$

Proposition

L'estimateur empirique $\bar{X}$ de l'espérance d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, calculé à partir d'un échantillon i.i.d. de cette loi, est gaussien :

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Loi de $\bar{X}$



n= 5
n= 10
n= 30
n= 100
dnorm(x,mu0,sigma0/sqrt(n))
0
5
10
15
5.8
6.0
6.2
6.4
6.6
6.8
x

Applications

- Si on considère que la mesure de pollution d'un véhicule de marque donnée suit $X_1 \sim \mathcal{N}(c_0, \sigma_0^2)$, alors

$$\begin{aligned}\mathbb{P}(\bar{X} \geq 6.43) &= \mathbb{P}\left(\frac{\bar{X} - c_0}{\sigma_0/\sqrt{n}} \geq \frac{6.43 - c_0}{\sigma_0/\sqrt{n}}\right) \\ &= 1 - \mathrm{F}_{\mathcal{N}}\left(\underbrace{\frac{6.43 - c_0}{\sigma_0/\sqrt{n}}}_{2.87}\right) \simeq 0.002\end{aligned}$$

- On peut aussi se demander quelle valeur q de consommation moyenne sur 30 véhicules est dépassée avec une probabilité donnée (par ex, $\alpha = 5\%$)

$$\mathbb{P}(\bar{X} \geq q) = 1 - \mathrm{F}_{\mathcal{N}}\left(\frac{q - c_0}{\sigma_0/\sqrt{n}}\right) = 0.05$$

Soit $q = c_0 + \underbrace{\mathrm{F}_{\mathcal{N}}^{-1}(0.95)}_{\text{quantile d'ordre 95\% de } \mathcal{N}(0,1)} \frac{\sigma_0}{\sqrt{n}} = 6.38$

Vous avez dit quantile ?

Définition

Le quantile (fractile) q_α d'ordre α d'une loi de fonction de répartition F, est défini par

$$q_\alpha = \inf\{x; \mathrm{F}(x) \geq \alpha\}$$

- Si F est continue et strictement croissante,

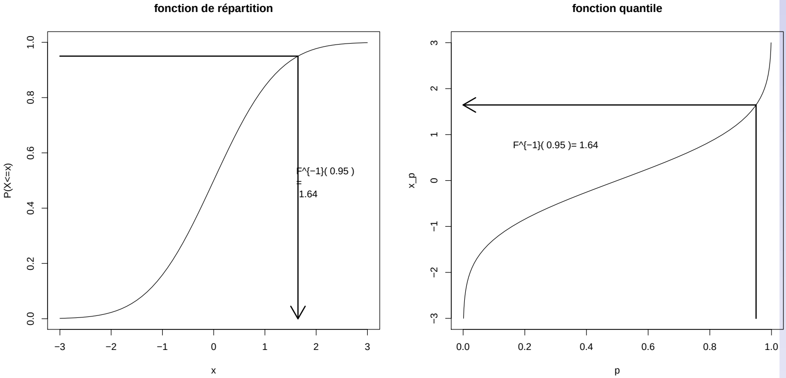
$$q_\alpha = \mathrm{F}^{-1}(\alpha)$$

- Si la loi est discrète, et α entre deux marches, on convient de faire une interpolation linéaire

Dans tous les cas, on notera F^{-1} la fonction quantile, inverse généralisé de F de $[0; 1]$ sur le domaine de définition de X.

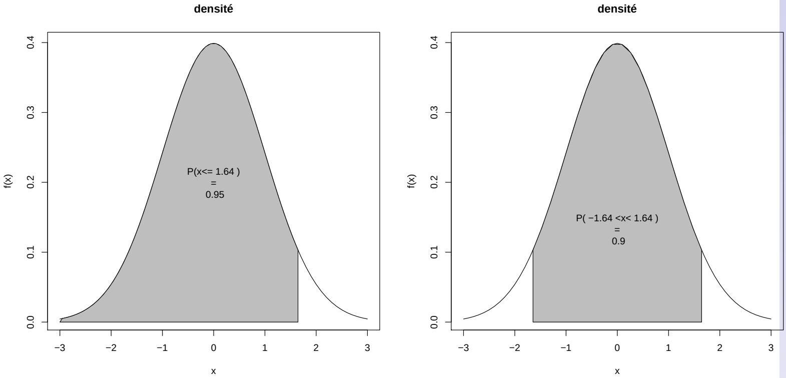
Fonction de répartition et fonction quantile

$X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$



α	0.9	0.95	0.975	0.99	0.999
q_α	1.28	1.64	1.96	2.33	3.09

Quantiles $X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$



Si la densité est symétrique, $q_\alpha = -q_{1-\alpha}$

Loi du Khi-deux

Définition (loi du Khi-deux)

Soit Z un vecteur gaussien centré réduit de dimension n. La loi de la somme du carré de ses composantes est la loi du Khi-deux (centré) à n degrés de liberté

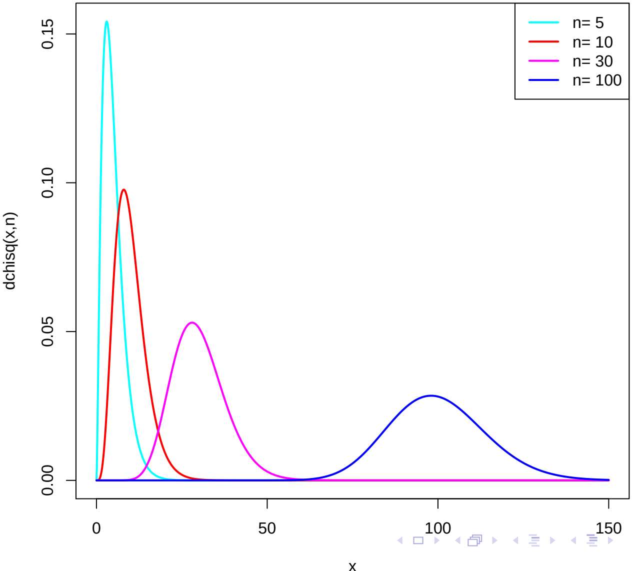
$$K_n = \sum_i Z_i^2 \sim \chi^2(n); \quad \psi_{K_n}(t) = \mathbb{E}(e^{tK_n}) = \frac{1}{(1-2t)^{n/2}}$$

$\mathbb{E}(K_n) = n; \quad \mathrm{Var}(K_n) = 2n$

Loi du Khi-2 décentrée : Si $Y \sim \mathcal{N}_n(\mu, Id_n)$, alors

$$||Y||^2 \sim \chi^2(n, ||\mu||^2)$$

Loi du $\chi^2(n)$



Loi de l'estimateur de la variance à espérance connue

Proposition

Soit $V_n^* = \frac{1}{n}\sum_i (X_i - \mu)^2$, l'estimateur empirique de la variance d'un échantillon i.i.d. X de loi $\mathcal{N}(\mu, \sigma^2)$, μ connue :

$$n\frac{V_n^*}{\sigma^2} = \sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Loi de la somme des carrés résiduels

Proposition

Si X un n-échantillon gaussien de variance σ^2, alors

$$\sum_{i}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

On en déduit, pour $S_n^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ et $\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$.

$$n\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1); \;\; (n-1)\frac{\widehat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1)$$

Preuve : Cochran

Rem : Dans le cas iid général, $\bar{X}$ et S_n^2 sont tels que

$$cov(\bar{X}, S_n^2) = \frac{n-1}{n^2}\mathbb{E}((X_1 - \mu)^3)$$

Projection d'un vecteur gaussien

Théorème (Cochran)

Si $Y \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$, *et si* $E_1 \oplus \ldots \oplus E_r = \mathbb{R}^n$ *est une décomposition de* $\mathbb{R}^n$ *en* r *sous-espaces orthogonaux, alors les projections orthogonales* $\Pi_1(Y), \ldots, \Pi_r(Y)$ *sur ces sous-espaces sont des vecteurs gaussiens indépendants tels que, pour tout* $j = 1, \ldots, r$

$$||\Pi_j(Y)||^2 \sim \sigma^2 \chi^2(d_j = Dim(E_j), \mu_j = ||\Pi_j(\mu)/\sigma||^2).$$

Preuve Th. de Cochran

- $Z = Y/\sigma$
- Pour tt j, soit $(e_{j_1}, \dots, e_{j_{d_j}})$ une base orthonormée de E_j

$$\Pi_j Z = \sum_{k=1}^{d_j} < e_{jk}, Z > e_{jk}$$

- Soit U matrice de passage $UU' = Id_n$. On a $UZ \sim \mathcal{N}_n(U\mu/\sigma, Id_n)$. Les variables $e'_{jk}Z$ sont indépendantes quand j et k varient. Donc $\Pi_1(Z), \dots, \Pi_r(Z)$ sont indépendantes
- Pour un sous-espace E_j, et pour $k = 1, \dots, k_j$

$$e'_{jk}Z \sim \mathcal{N}(e'_{jk}\mu, e'_{jk}e_{jk} = 1)$$

- d'où , avec $\mu_j = ||\Pi_j \mu/\sigma||^2 = \sum_{k=1}^{d_j} (e'_{jk}\mu/\sigma)^2$

$$||\Pi_j(Z)||^2 = \sum_{k=1}^{d_j} ||e'_{jk}Z||^2 \sim \chi^2(d_j, \mu_j),$$

Loi de Student (Gosset)

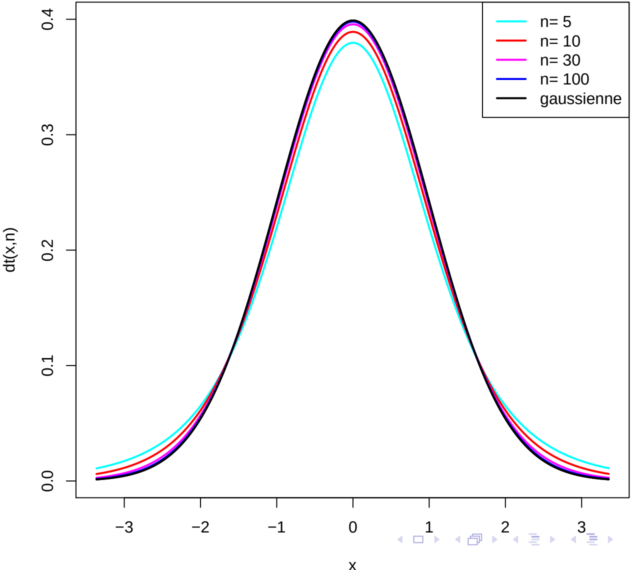
Définition (Loi de Student)

Soit deux variables Z et K indépendantes telles que $Z \sim \mathcal{N}(0,1)$ et $K \sim \chi^2(p)$. Alors, la v.a.

$$T = \frac{Z}{\sqrt{\frac{K}{p}}} \sim \mathcal{T}(p)$$

suit une loi appelée loi de Student à p degrés de liberté.

Loi de Student



Statistique de Student

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Proposition

Si $X_1, \ldots, X_n$ est un n-échantillon gaussien de loi $\mathcal{N}(\mu, \sigma^2)$,

$$\sqrt{n}\frac{\bar{X} - \mu}{\widehat{\sigma}_n} \sim \mathcal{T}(n-1) \quad \textit{avec } \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Application : $\mathbb{P}(\bar{X} > 6,43) = 1 - \mathrm{F}_{\mathcal{T}}(\underbrace{\frac{6.43 - c_0}{0.25/\sqrt{n}}}_{2.41}, n-1) \simeq 0.011$

Loi de Fisher

Définition (Loi de Fisher)

Soit deux variables K_1 et K_2 indépendantes telles que $K_1 \sim \chi^2(n_1)$ et $K_2 \sim \chi^2(n_2)$. Alors, la v.a.

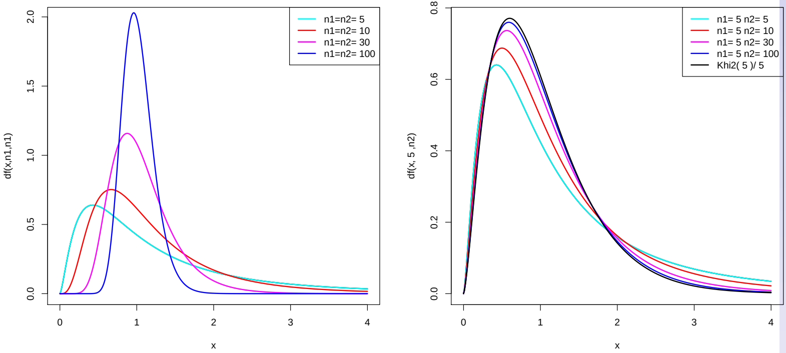
$$F = \frac{K_1/n_1}{K_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

suit une loi appelée loi de Fisher à (n_1, n_2) degrés de liberté.

Proposition

$\mathbb{E}(F)$ *existe pour* $n_2 \geq 2$ *et vaut* $\mathbb{E}(F) = \frac{n_2}{n_2-2}$. $\mathrm{Var}(F)$ *existe pour* $n_2 \geq 5$ *et vaut* $\mathrm{Var}(F) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$

Loi de Fisher



Application de la loi de Fisher

Proposition (Loi du rapport des estimateurs de variance)

Soient deux échantillons gaussiens indépendants de taille n_1 et n_2, de même variance σ^2, et soient $\widehat{\sigma}_1^2$ et $\widehat{\sigma}_2^2$ les estimateurs non biaisés de la variance σ^2 dans chacun des deux échantillons. Alors, la v.a.

$$\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1)$$

Approximation gaussienne

Si la loi mère n'est pas gaussienne, et si la loi de $\bar{X}$ est difficile à identifier, on peut utiliser des approximations gaussiennes pour des échantillons suffisamment grands.

Théorème (de limite centrale)

Soit $\{X_n\}$ une suite de variables aléatoires i.i.d. admettant une espérance μ et une variance σ^2 finie. Alors, la suite des variables $\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}$ converge en loi vers la v.a. $\mathcal{N}(0,1)$ quand $n \to \infty$

$$\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Utilisation de Slutsky

Proposition

Si $X_1, \ldots, X_n$ est un n-échantillon de loi d'espérance μ et de variance σ^2 finie,

$$\sqrt{n}\frac{\bar{X}-\mu}{\widehat{\sigma}_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \textit{avec } \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\sqrt{n}\frac{\bar{X}-\mu}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \textit{avec } S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Normalité asymptotique

Définition

Soit un estimateur $\widehat{\nu}_n$ *de* $\nu \in \mathbb{R}^p$*. S'il existe une v.a.* V_n *telle que*

$$V_n^{-1/2}(\widehat{\nu}_n - \nu) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Id_p)$$

on dit que l'estimateur est asymptotiquement normal. Si $nV_n \to V_0$ *où* $V_0 > 0$ *est finie, on dit que la vitesse de l'estimateur est en* $\sqrt{n}$

$\hookrightarrow$ Un estimateur est d'autant meilleur que sa vitesse de convergence est rapide et sa loi limite concentrée autour de 0.

Delta-méthode

Proposition

Si h est une fonction différentiable de $\nu \in \mathbb{R}^p$ et $\widehat{\nu}_n$ un estimateur asymptotiquement normal, alors $h(\widehat{\nu}_n)$ est un estimateur asymptotiquement normal de $h(\nu)$

$$(D_\nu V_n(\nu) D'_\nu)^{-1/2}(h(\widehat{\nu}_n) - h(\nu)) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Id_p)$$

avec $D_\nu = \begin{pmatrix} \partial h(\nu)/\partial \nu_1 & \dots & \partial h(\nu)/\partial \nu_p \end{pmatrix}$