

Statistiques et probabilités

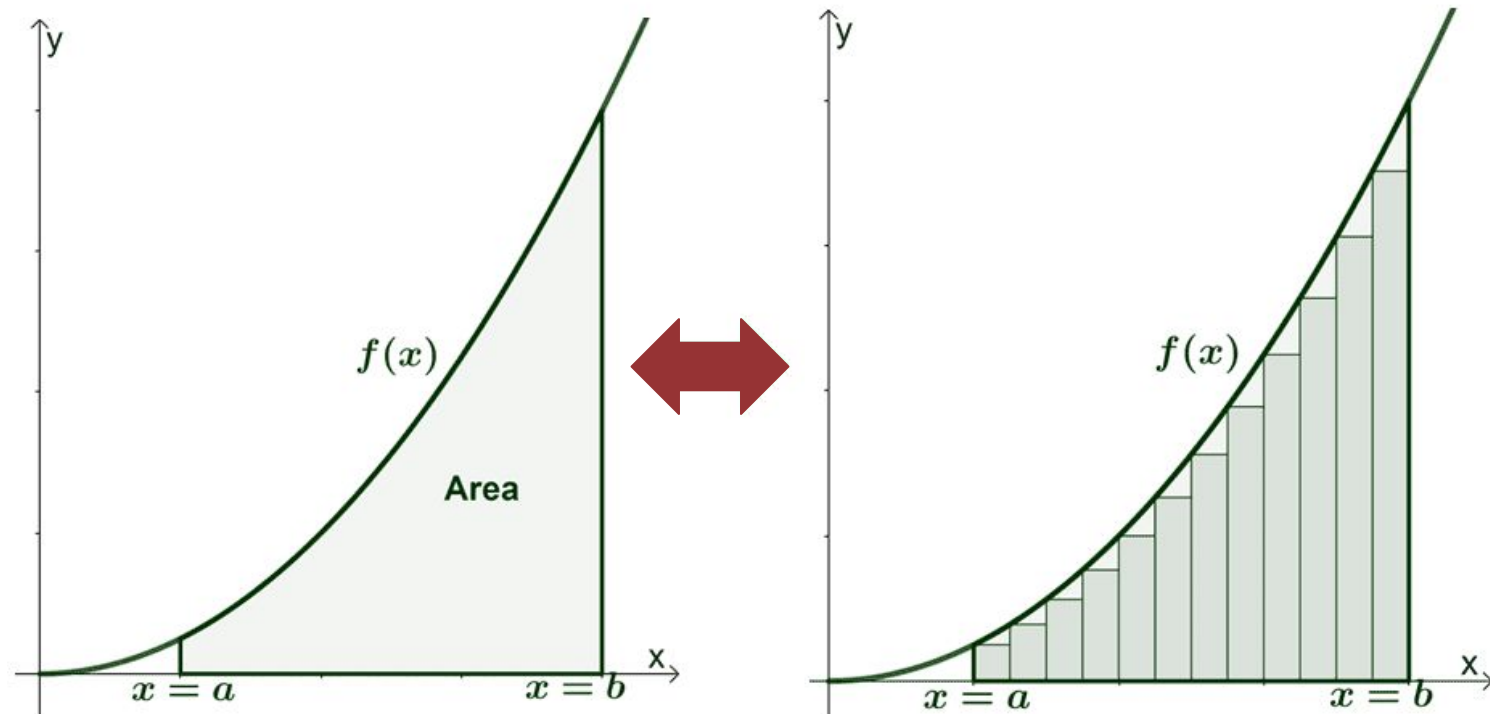
Cours n°4

Guillaume Postic

Université Paris-Saclay, Univ. Evry
Département informatique

Master 1 MIAE - 2023/2024

Rappels



Rappels

- La somme de Riemann S de f sur $[a, b]$ est :

$$S = \sum_{i=1}^n f(x_i^*) \Delta x_i$$

- Si le **pas de la subdivision** tend vers zéro
(i.e. si $n \rightarrow +\infty$), alors la somme converge vers :

$$\int_a^b f(t) \, dt$$

Variable aléatoire continue

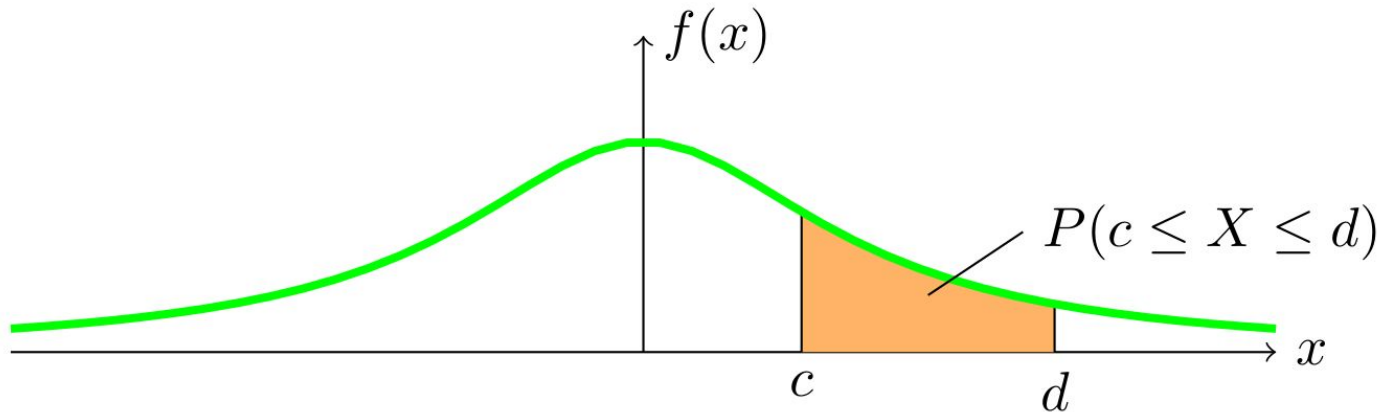
- Intervalle de valeurs continu :
 - e.g. $[0, 1]$, $[a, b]$, $[0, \infty)$, $(-\infty, \infty)$
- Fonction de densité ou **densité de probabilité**

$$f(x) \geq 0; \quad P(c \leq x \leq d) = \int_c^d f(x) dx$$

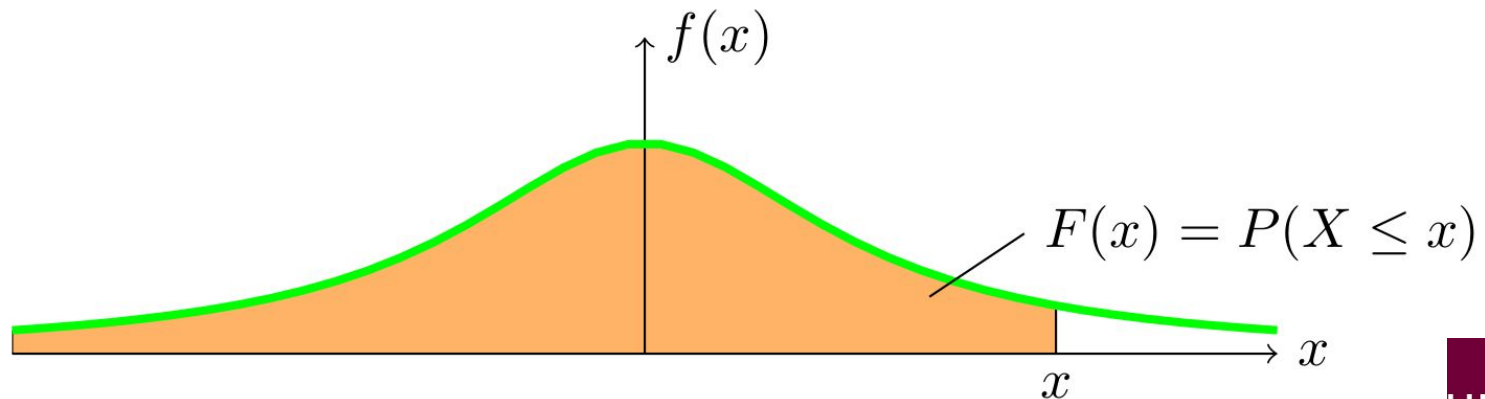
- Fonction de répartition

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Variable aléatoire continue



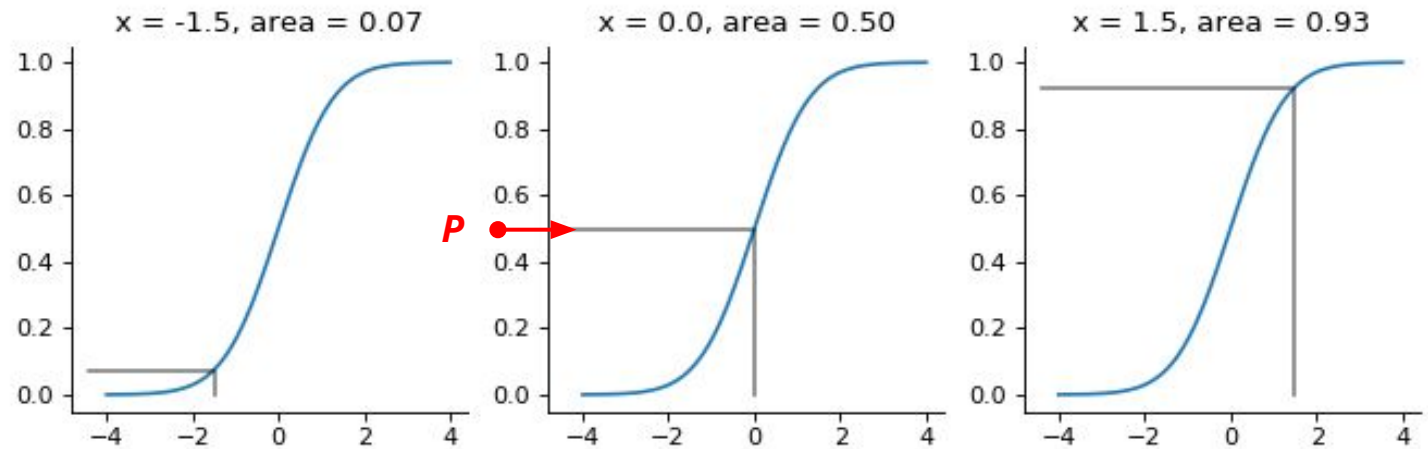
Fonction de densité et probabilité



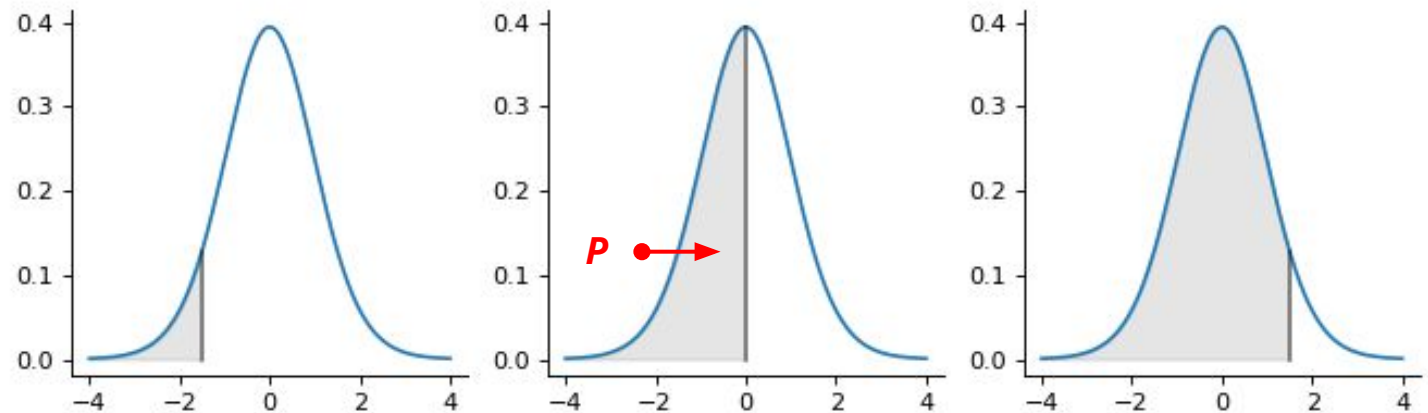
Fonction de densité et fonction de répartition

Fonction de répartition

Fonction de répartition
 $F(x)$



Fonction de densité
 $f(x)$



Propriétés de la fonction de répartition

NB : également valables pour les variables aléatoires discrètes

- (Définition) $F(x) = P(X \leq x)$
- $0 \leq F(x) \leq 1$
- Jamais décroissante
- $\lim_{x \rightarrow -\infty} F(x) = 0$
- $\lim_{x \rightarrow +\infty} F(x) = 1$
- $P(c < X \leq d) = F(d) - F(c)$
- $F'(x) = f(x)$

Espérance

X continue sur $[a, b]$, avec la fonction de densité $f(x)$:

$$E(X) = \int_a^b xf(x) dx$$

X discrète, de valeurs x_1, \dots, x_n , avec la fonction de masse $p(x_i)$:

$$E(X) = \sum_{i=1}^n x_i p(x_i)$$

Variance et écart-type

Pour toute variable aléatoire X de moyenne μ

$$\text{Var}(X) = E((X - \mu)^2), \quad \sigma = \sqrt{\text{Var}(X)}$$

X continue sur $[a, b]$, avec la fonction de densité $f(x)$:

$$\text{Var}(X) = \int_a^b (x - \mu)^2 f(x) dx.$$

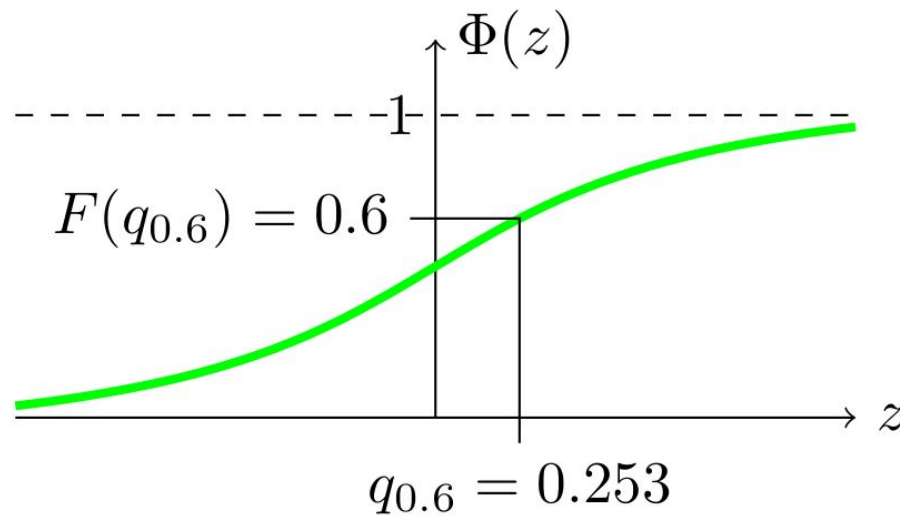
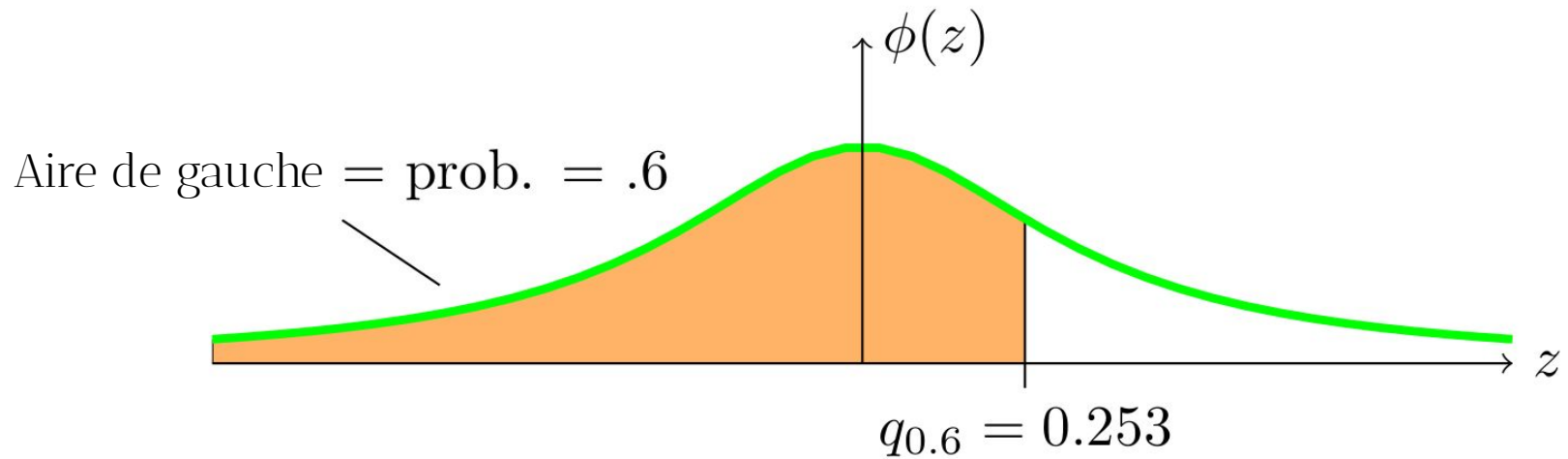
X discrète, de valeurs x_1, \dots, x_n , avec la fonction de masse $p(x_i)$:

$$\text{Var}(X) = \sum_{i=1}^n (x_i - \mu)^2 p(x_i).$$

Quantiles (1)

- Les quantiles sont les valeurs qui divisent un jeu de données en intervalles de probabilités égales.
 - Il y a donc un quantile de moins que le nombre de groupes créés. Par exemple, **les quartiles sont les trois quantiles** qui divisent un ensemble de données en quatre groupes de même probabilité.
- La **médiane** (second quartile) sépare le jeu de données en deux groupes de même probabilité.
 - C'est un **indicateur de tendance centrale**

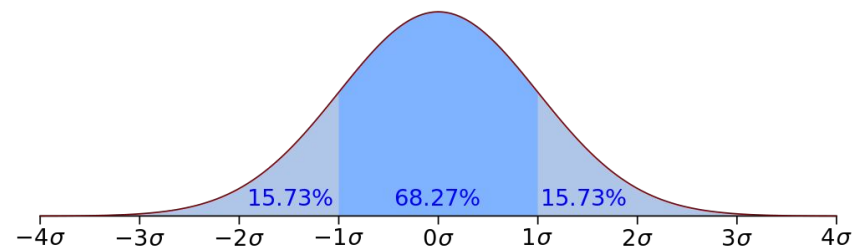
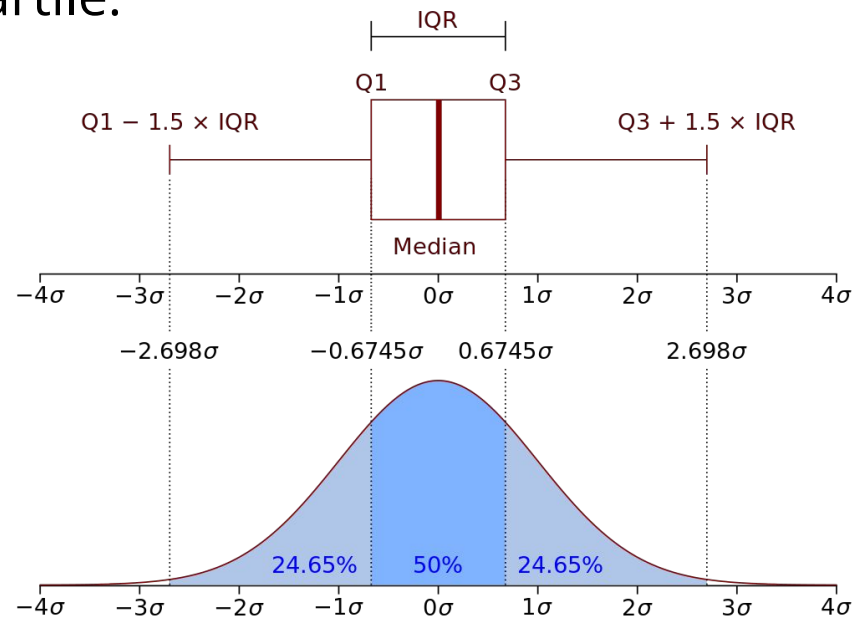
Quantiles (2)



Écart interquartile

Indicateur de dispersion qui s'obtient en faisant la différence entre le troisième et le premier quartile.

Représentation en
« boîte à moustaches »
(*box plot*)



Distributions continues (1)

Loi exponentielle

Loi de probabilité continue qui **modélise le temps x** s'écoulant entre deux occurrences d'un **processus de Poisson** de paramètre **λ** , nombre moyen d'occurrences dans un **intervalle de temps fixé**.

Fonction de densité :

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Fonction de répartition :

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

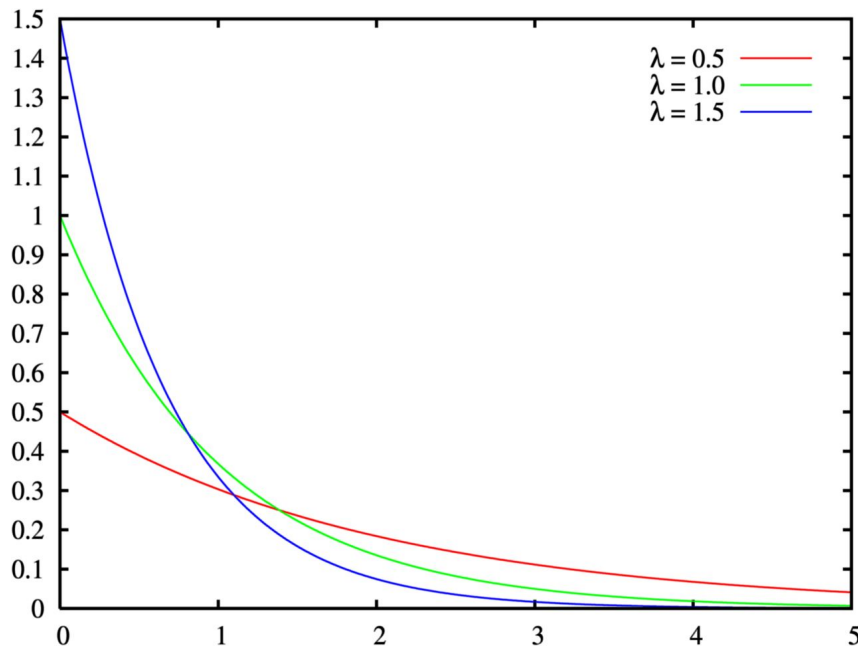
Espérance : $1/\lambda$

Variance : $1/\lambda^2$

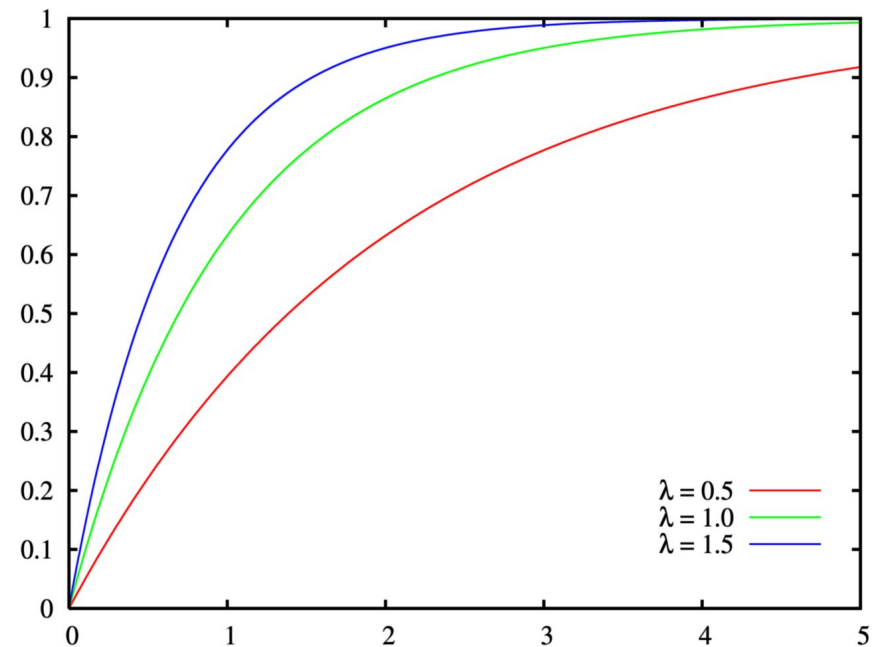
Distributions continues (1)

Loi exponentielle

$$X \sim \exp(\lambda)$$



Fonction de densité



Fonction de répartition

Distributions continues (2)

Lois uniformes continues

Les lois uniformes continues forment une famille de lois de probabilité à densité caractérisées par la propriété suivante : tous les intervalles de même longueur inclus dans le support de la loi ont la même probabilité.

Fonction de densité :

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{pour } a \leq x \leq b, \\ 0 & \text{sinon.} \end{cases}$$

Espérance : $(a+b)/2$

Variance : $(b-a)^2/12$

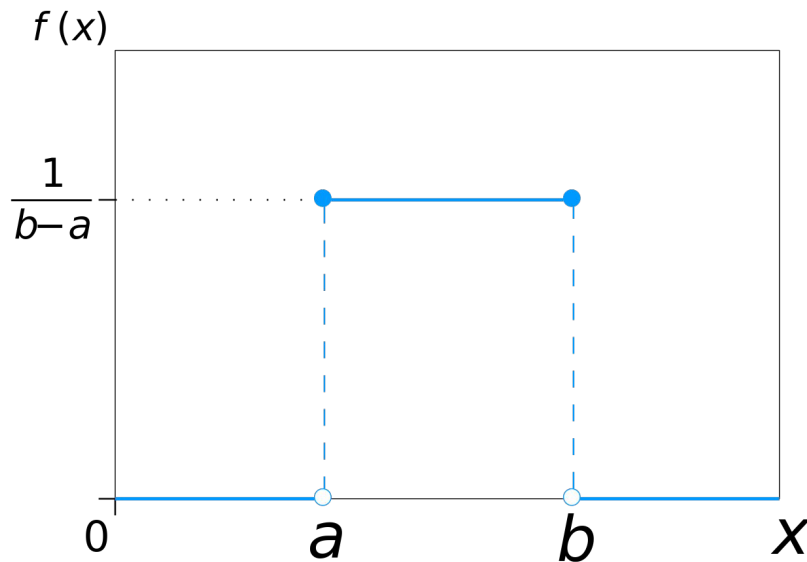
Fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{pour } x < a \\ \frac{x-a}{b-a} & \text{pour } a \leq x < b \\ 1 & \text{pour } x \geq b \end{cases}$$

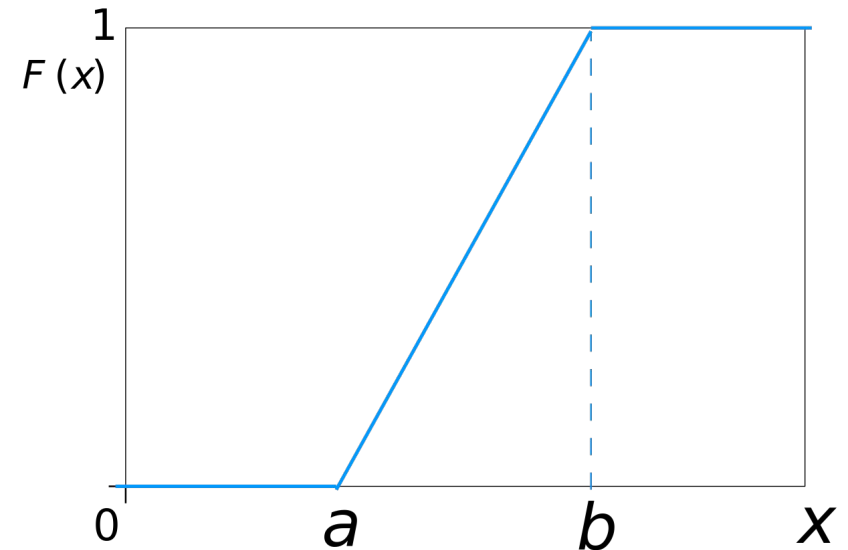
Distributions continues (2)

Lois uniformes continues

$$X \sim U(a, b)$$



Fonction de densité



Fonction de répartition

Distributions continues (3)

Loi normale

- Également appelée loi gaussienne, loi de Gauss ou loi de Laplace-Gauss
- Utilisées pour modéliser des **phénomènes naturels** issus de plusieurs événements aléatoires

Fonction de densité :

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

Fonction de répartition :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

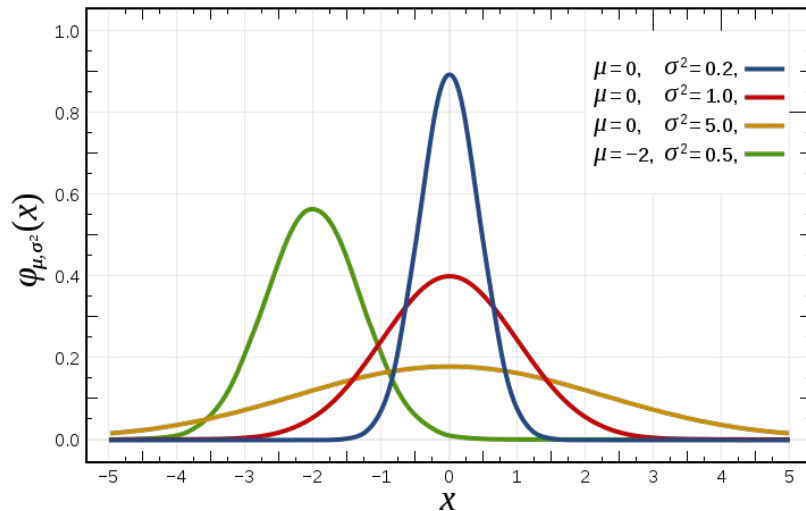
Espérance : μ

Variance : σ^2

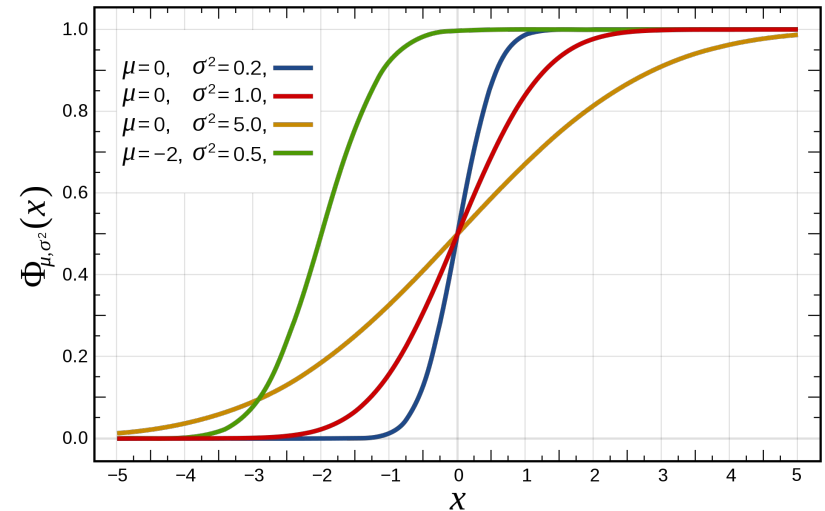
Distributions continues (3)

Loi normale

$$X \sim N(\mu, \sigma^2)$$



Fonction de densité

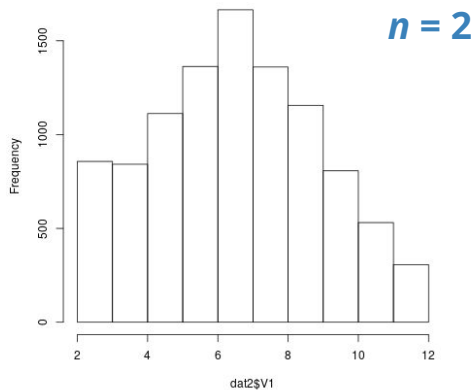


Fonction de répartition

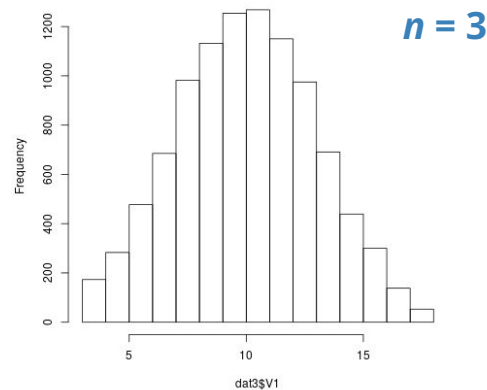
Simulations

On lance **10 000 fois** n dés équilibrés à 6 faces ; à chaque lancer, on calcule **la somme S des valeurs des n dés.**

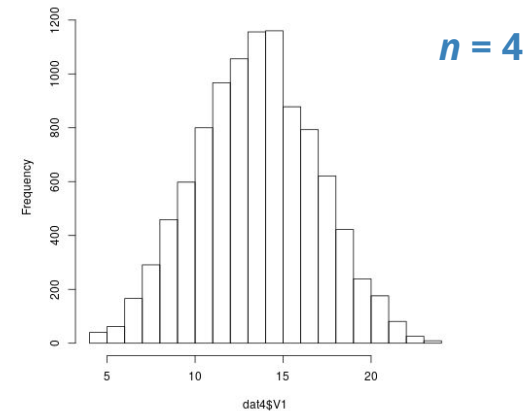
Histogram of dat2\$V1



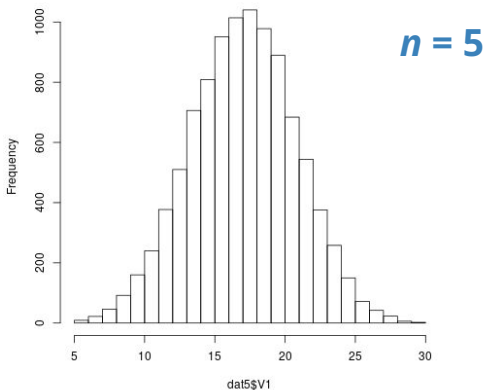
Histogram of dat3\$V1



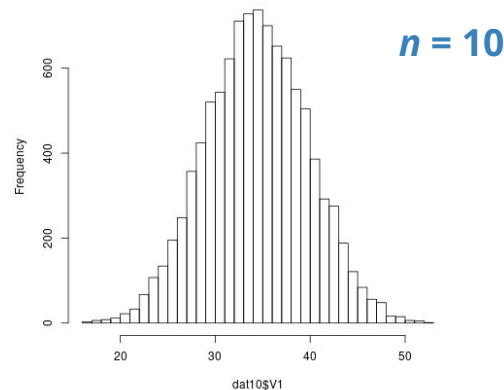
Histogram of dat4\$V1



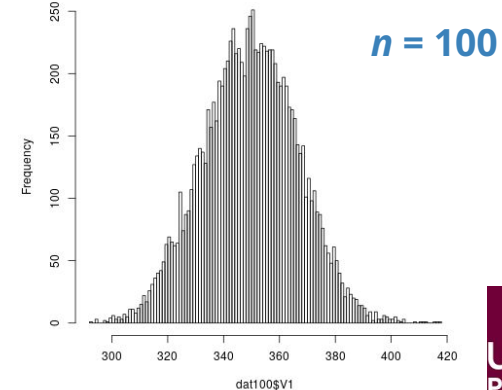
Histogram of dat5\$V1



Histogram of dat10\$V1



Histogram of dat100\$V1



Théorème limite central

Soit n **variables aléatoires « i. i. d. »**,

- Indépendantes
- Identiquement distribuées

Soit S , la somme de leurs valeurs.

Quand $n \rightarrow +\infty$, la loi de probabilité de S **converge** vers une loi normale de paramètres $E[S]$, $\text{Var}(S)$.

En pratique, **l'approximation** en une loi normale est souvent faite à partir de $n > 20$ ou 30.

Loi des grands nombres

- Note : un autre « théorème limite »
- La **moyenne empirique (calculée sur les valeurs d'un échantillon)** converge vers l'**espérance (moyenne de la population)** lorsque la taille de l'échantillon augmente.

Soit \bar{X}_n , moyenne d'un échantillon de taille n .

Soit μ , moyenne de la population.

- **Loi faible** (convergence en probabilité)

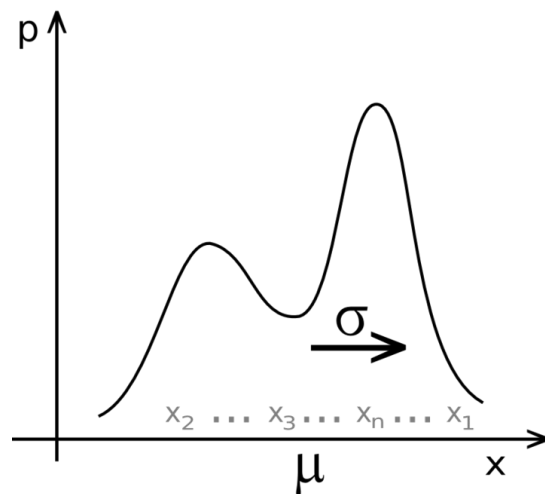
$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

- **Loi forte**

$$P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$$

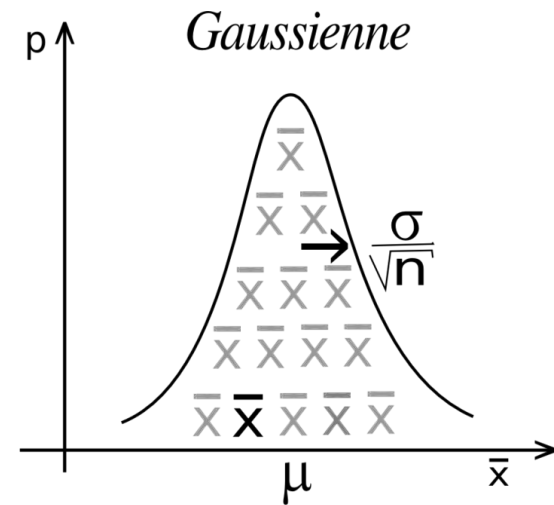
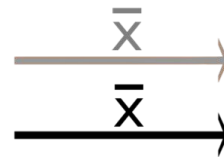
Statistiques inférentielles

Le TLC **s'applique donc à la distribution des moyennes d'échantillons** de tailles égales **et tirés indépendamment**.
(ces moyennes étant des sommes, toutes divisées par la même valeur)



distribution
de la
population

échantillons
de taille n



distribution
d'échantillonnage

Estimateur de l'écart-type : $S \sim N(n\mu, n\sigma^2)$

Erreur type de la moyenne

Standard error of the mean (SEM) vs *Standard deviation (SD)*

L'erreur type de la moyenne est une mesure de la dispersion des moyennes des tirages autour de la moyenne de la population.

Si l'écart type σ de la population est connu :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sinon, on utilise un estimateur s de σ (cf. biais et correction de Bessel) :

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

Remarques

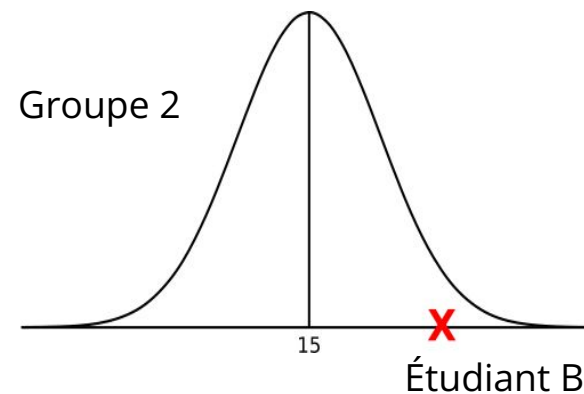
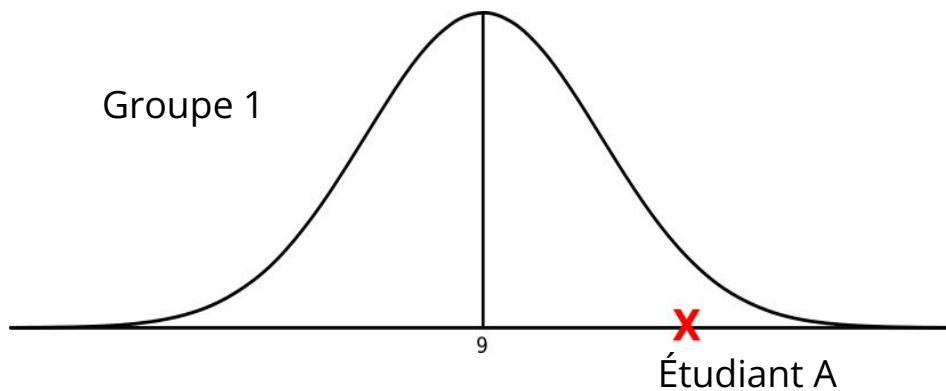
- Le TLC ne s'applique pas si la variance est
 - indéfinie (p. ex. loi de Cauchy)
 - infinie (p. ex. loi de Pareto)
- La loi binomiale (discrète) peut être approximée par une loi normale (continue)
 - Il est également possible d'approximer une loi continue par une loi discrète (cf. discrétisation et histogrammes)
- La **convolution** de n densités tend vers la densité normale quand $n \rightarrow +\infty$ (cf. « théorème local limite »)
- Généralisation du TLC : suppression de l'hypothèse que les variables sont de même loi (cf. conditions de Liapounov et Lindeberg)

Qui est le meilleur ?



Deux étudiants veulent savoir qui est le meilleur, en comparant leurs notes obtenues à l'UE de statistiques. Ils appartiennent à deux groupes de TD différents, chacun noté par un enseignant différent.

- L'étudiant A a eu 15/20, dans le groupe 1, où la moyenne est de 9 et l'écart-type est de 5.
- L'étudiant B a eu 19/20, dans le groupe 2, où la moyenne est de 15 et l'écart-type est de 3.



Normalisation

- Redimensionner les variables quantitatives pour qu'elles soient comparables sur une **échelle commune**

- *Min-max scaling*

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- *Mean normalization*

$$x' = \frac{x - \bar{x}}{\max(x) - \min(x)}$$

- Transformation en (per)centiles

Standardisation

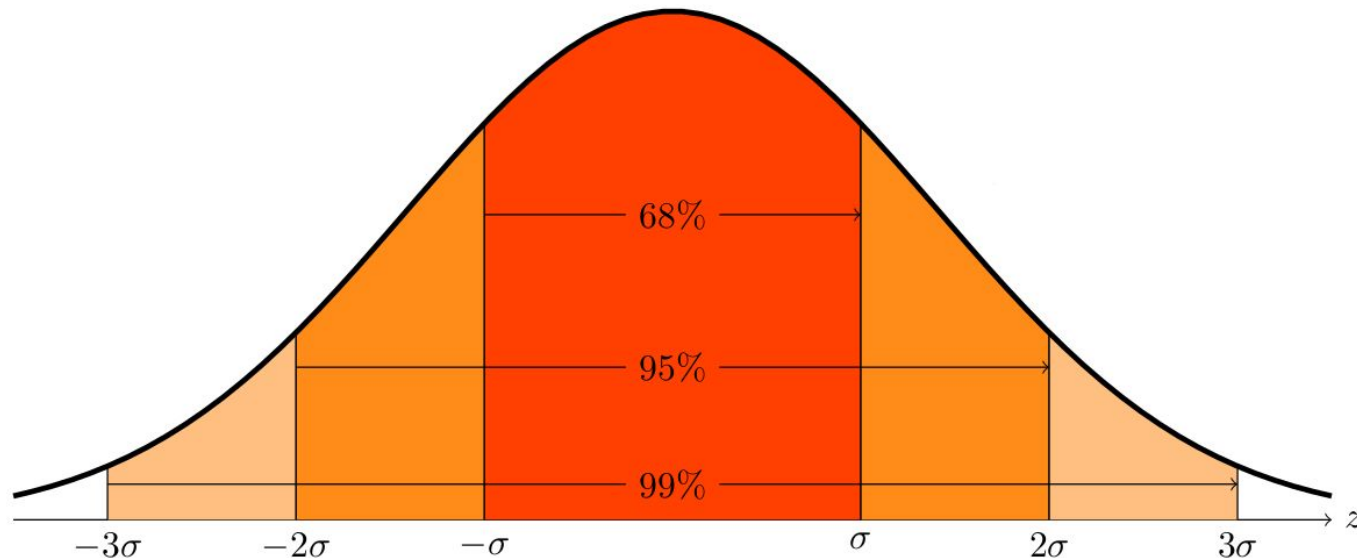
- Ou « normalisation **standard** »
- Transformation en une variable z
 - **centrée** : soustraction de la moyenne μ
 - **réduite** : division par l'écart type σ (**standard deviation**)

$$z = \frac{x - \mu}{\sigma}$$

- Ainsi, la distribution de z aura pour paramètres : $\mu = 0$ et $\sigma = 1$
- La variable standardisée est appelée **z-score** ou cote Z
- **Un z-score égal à a signifie que la donnée x s'éloigne de la moyenne de a écarts-types** (donc $Z < 0$ si $x < \mu$)
- Si $X \sim N(\mu, \sigma)$, alors $Z \sim N(0, 1)$

Loi normale centrée réduite

Lecture de la loi normale standardisée ou **fonction de distribution Z** :



1. $P(-1 < Z < 1)$ is

- (a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

2. $P(Z > 2)$

- (a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

Qui est le meilleur ?



Réponse : Les notes centrées réduites sont directement comparables. L'étudiant A a une note standardisée de 1,20, inférieure à celle de l'étudiant B, 1,33.

Cas d'une loi non-connue

Soit $Z = (x - \mu) / \sigma > 2$

- si $X \sim N(\mu, \sigma)$, alors $p(X > x) = 2,5\%$ (environ)
- si $X \sim$ loi non-connue (de moyenne μ et variance σ^2)
 - **l'inégalité de Bienaymé-Tchebychev** : donne la probabilité « dans le pire des cas »
 - $p(X > x) < 1/Z^2 = 25\%$
 - $p(X < x) < 1/Z^2 = 25\%$

Distributions continues (4)

Loi du χ^2

- Prononcé « khi carré » ou « khi-deux »
- Si Z_1, Z_2, \dots, Z_k sont des VA suivant la **loi normale standardisée**, alors la **somme de leurs carrés**

$$X = Z_1^2 + Z_2^2 + \dots + Z_k^2$$

suit une loi du χ^2 avec **k degrés de liberté**

Fonction de densité :

$$\frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Fonction de répartition :

$$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$$

Espérance : k

Variance : $2k$

Distributions continues (4)

Loi du χ^2

$$X \sim \chi^2(k)$$

