Statistique Univariée

Marie Beurton-Aimar

Plan

Objectifs

- Controler/Vérifier les informations.
- Interagir avec les statisticiens.

- Maitriser le langage de base en statistique
- Organiser les données sous forme de tableaux, de graphes.
- Calculer les paramètres de statistique descriptive
- Réaliser l'analyse statistique descriptive avec le logiciel R.

Le langage en statistique

Variable

- Notion fondamentale, c'est la caractéristique observée/mesurée.

Variable quantitative

- Une variable quantitative est mesurable.
- Quantitative continue : un nombre infini de valeurs.
- Quantitative discrète: un nomber fini (dénombrable) de valeurs .
- Exemple la taille des souris est continue, le nombre de souris dans l'expérience est discret.

Le langage en statistique

Variable qualitative

- Une variable qualitative est discrète.
- Elle peut prendre plusieurs valeurs appelées **modalités**.
- Ordinale si ses modalités présentent une hiérarchie.
- Sinon variable nominale.
- Exemple : la variable sexe des souris vaut F ou M - c'est une variable nominale.
- Exemple : la couleur du pelage, plus ou moins foncé.

Le langage en statistique

Compléments

- X dépend de Y si la distribution de X dépend de celle de Y.
- X - Variable **dépendante** - expliquée.
- Y - Variable **indépendante** explicative.
- Exemple : expliquer la croissance bactérienne par la température d'incubation.

Le langage en statistique

Observation

- L'**individu** ou observation: c'est l'unité statistique sur laquelle les variables observées seront mesurées.
- La **population** : ensemble dont sont issus les individus observés. En général, impossible à étudier dans sa totalité.
- L'**échantillon** : sous-ensemble de la population. Constitué par tirage aléatoire, si son effectif n est 30 on considère que l'on pourra s'affranchir des fluctuations d'échantillonage.

Le langage en statistique

Question posée

- L'analyse statistique est réalisée dans le cadre d'une **Hypothèse théorique** de travail (problématique posée).
- L'**hypothèse opérationnelle** détermine : la population à éteudier et les variables à prendre en compte : variable dépendante, variables indépendantes

Plan d'expérience

- Le plan d'expérience détermine les outils statistiques choisis : choix de l'effectif en fonction du nombre de facteurs, de leur nature, des variables étudiées ....

Statistique descriptive ou statistique exploratoire

- Permet d'organiser, de visualiser et de résumer l'information contenue dans de volumineux jeux de données.

Les tableaux

Tableau d'effectifs

- Pour un effectif n, on note n_i le nombre d'individus que présente la valeur x_i.
- Pour une vision plus claire de la distribution d'une variable quantitative continue, on regroupe par classes les valeurs voisines.
- Une classe est définie par :
  - ses bornes incluses ou exclues $[e_i, e_{i+1}[$
  - son amplitude $h = e_{i+1} - e_i$
  - son centre $C_i = e_i + \frac{e_{i+1} - e_i}{2}$

Les tableaux

Tableau de fréquences

- Principe : on peut utiliser les fréquences à la place des effectifs.
- Pour chaque valeur x_i de la variable on note sa fréquence $f_i = n_i/n$.
- **Remarque** : on peut s'intéresser aux fréquences cumulées croissantes ou décroissantes.

Valeur de variable (X_i)	Effectif(n_i)	Fréquence (f_i)
X_1	n_1	f_1
X_2	n_2	f_2
.	.	.
.	.	.
X_n	n_n	f_s

- **Tableau de contingence** : vu au TD3.

Les tableaux

Utilisation

- Rangement des données dans un tableau.
- Classement des valeurs par ordre croissant (ou décroissant).
- Définition du nombre d'individus par valeur/modalité ou par classe.
- Calcul de fréquences.

Représentation graphique

Objectif

Visualiser les données collectées dans un tableau.

Les représentations utilisées

- Histogrammes.
- Diagrammes en boites.
- Diagrammes en batons.
- Diagrammes pour variables quantitatives.

Représentation graphique

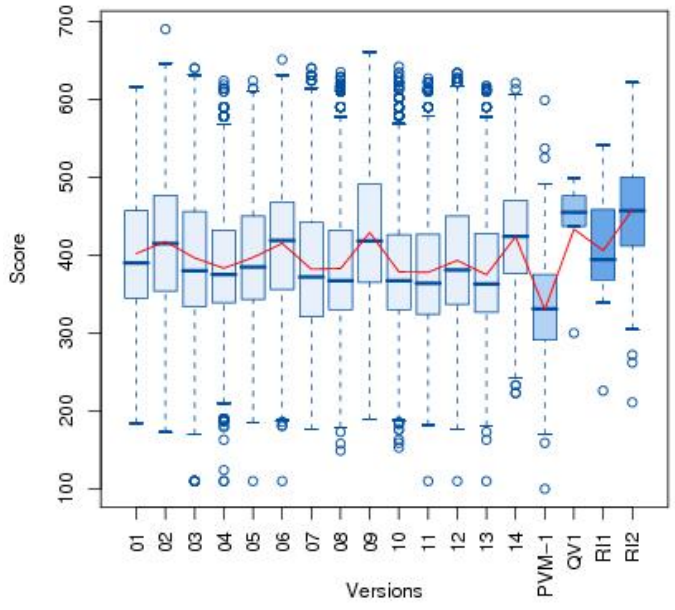
Histogramme

- Distribution des effectifs d'une variable quantitative continue.
- Aire des rectangles proportionnelle à l'effectif de chaque classe.
- Une ligne brisée, appelée le **polygone des fréquences** relie les centres de classes par des segments de droite.
- La courbe limite s'appelle **densité de probabilité**.

Représentation graphique

Diagramme en boites

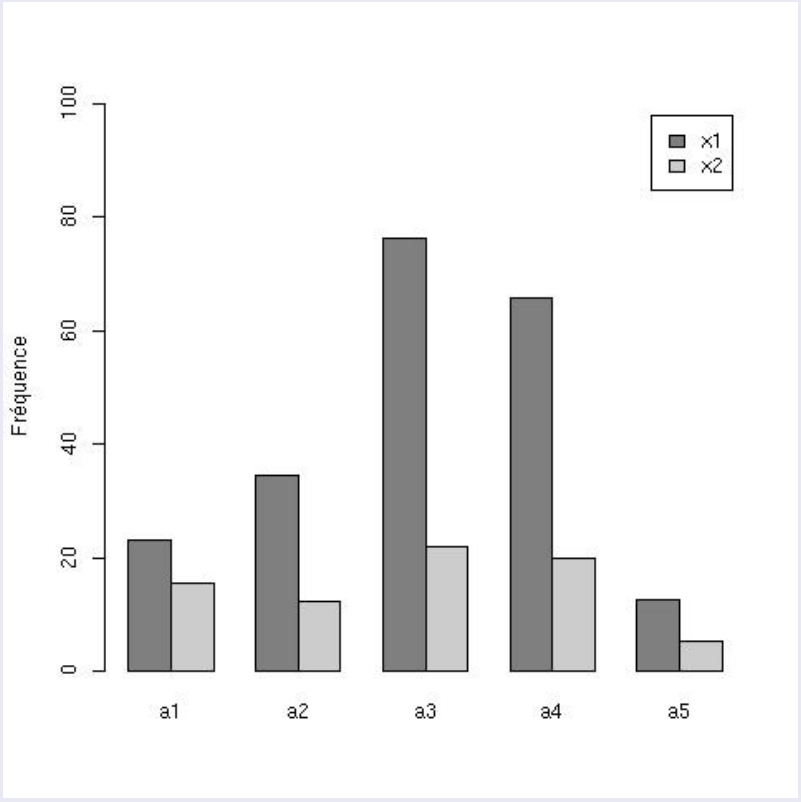
- Boite à moustaches : représentation horizontale
- Cotés de la boite : quartiles Q1 et Q3, le centre Q2
- Extensions (moustaches) : 1,5 fois l écart interquartile (Q3-Q1)
- En dehors des extensions : valeurs aberrantes (outliers)
- Comparaison de plusieurs séries statistiques de même nature



Représentation graphique

Diagramme en batons

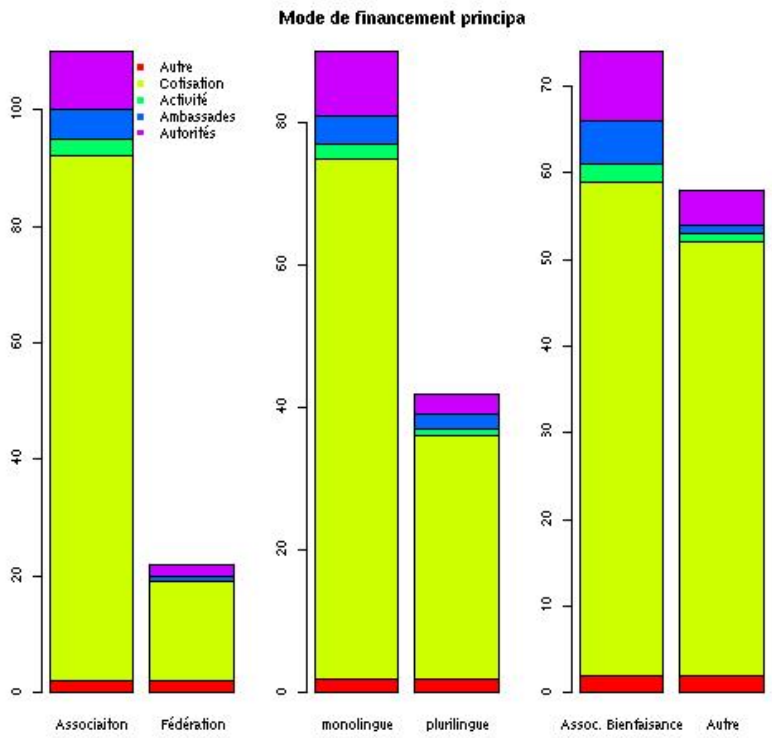
- Dédié aux variables quantitatives discrètes

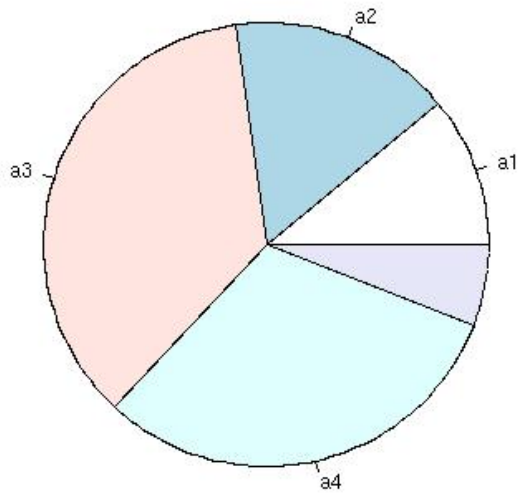


Représentation graphique

Diagramme pour les variables qualitatives

- Circulaires, à bandes ou en barres.
- Chaque modalité est représentée par un secteur ou un rectangle dont l'aire est proportionnelle à son effectif.





Paramètres de distribution

- Les paramètres de **position** :
  - **Moyenne arithmétique** : sensible aux valeurs extrèmes
  - **Médiane** : valeur qui partage l'effectif en deux.
  - **Mode** : valeur ayant l'effectif maximum
  - **Quartiles** : partage la série statistique ordonnée en 4 sous-ensembles d'effectifs identiques.

Paramètres de distribution

- Les paramètres de **dispersion** :
  - **Variance** : moyenne des carrés des écarts à la moyenne.
  - L'**écart type** (standard deviation) : racine carrée de la variance.
  - **Coefficient de variation** : rapport entre l'écart type et la moyenne d'une série - souvent exprimé en pourcentage.
  - **Ecart interquartile** (Q3-Q1) permet d'évaluer la dispertion des valeurs qui entourent la médiane.

Prise en main de R

Présentation

- R est un langage interprété.
- C'est une implémentation du langage S. Développé à l'origine par Robert Gentleman and Ross Ihaka. C'est un projet GNU maintenu principalement en Nouvelle Zélande par l'université d'Auckland.
- Utilisation relativement simple et intuitive.
- Utilisation souple :
  - Saisie interactive ou scripts
  - Création de données manuelle (nombreux outils) / chargement de données (fichier).

Premier Pas

Création et Modification de données

Exemples :

```
> x <-10
> x
[1] 10
> x<-15+12
> x
[1] 27
> x<-x+12
> x
[1] 39
```

Création et Modification de données

```
> x<-c(12,3,2,11,23,3,21)
> x
[1]12 3 2 11 23 3 21
> x[3]<-42
> x
[1] 12 3 42 11 23 3 21
> x<-x+12
> x
[1] 24 15 54 23 35 15 33
```

Génération de données

Opérateur **:**

```
> x<-1:6
> x
[1]1 2 3 4 5 6
> x<-6:1
> x
[1] 6 5 4 3 2 1
```

Génération de données

Opérateur **seq**

```
> x<-seq(from=1,to=10)
> x
[1] 1 2 3 4 5 6 7 8 9 10
> x<-seq(from=1,to=10,by=2)
> x
[1] 1 3 5 7 9
> x<-seq(length=21,from=0,to=1)
> x
[1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
[10] 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85
[19]0.90 0.95 1.00
```

Premier pas

Chargement de données

- Se placer dans le bon répertoire de travail : **setwd(“path”)**
- Importer des données : **read.table(“nomDuFichier”, header=TRUE/FALSE)**

Obtenir de l’aide

- Aide générale : **help.start()**
- Aide sur une fonction connue : **help(nomFonction)** ou **?(nomFonction)**.
- Exemple **?(cos)**

Distribution de probabilités

- Loi binomiale.
- Loi de Poisson.
- Loi de Gauss ou normale.
- Distribution multimodale.
- Distribution uniforme.

Distribution de probabilités

Loi binomiale

- Pour les variables discretes
- Basée sur l'expérience de Bernouilli.

Expérience de Bernouilli

- Expérience aléatoire discrète avec 2 résultats possibles : Succès ou échec.
- $p + q = 1$, p probabilité de succès, q probabilité d'échec.
- Expériences indépendantes : répétées à l'infini.
- Chaine de Bernouilli : épreuve de Bernouilli répétée n fois.
- Coefficient binomial : nombre de chaines de Bernouilli (n).
- La variable aléatoire X comptabilisant le nombre de succès sur les n épreuves suit une loi binomiale de paramètre n et p notée $B(n, p)$.

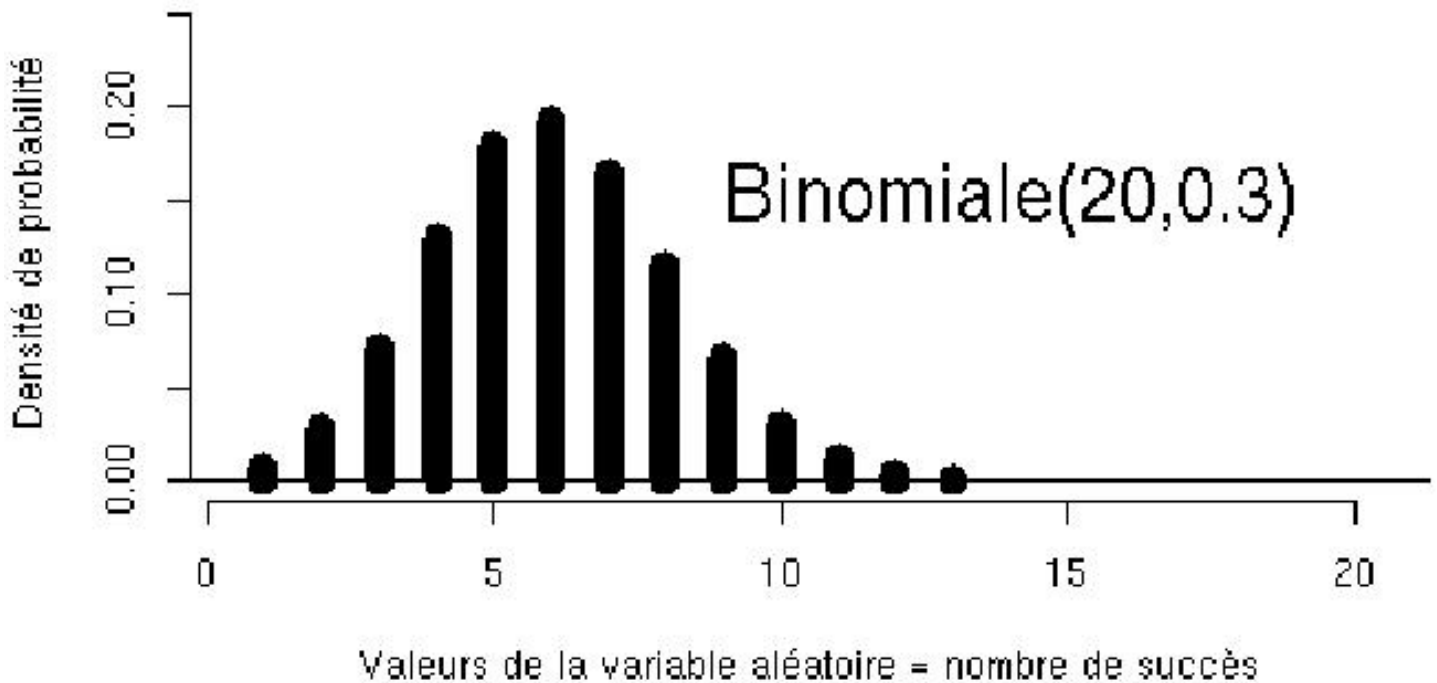
Loi binomiale

Commande R

- rbinom(N,n,p)

Moments d'une loi binomiale $B(n, p)$

- $E(X) = n \times p$
- $V(X) = n \times p \times q$



Distribution de probabilités

Loi de Poisson

- Variable aléatoire discrète.
- Modèle probabiliste des situations de flux d'événements successifs et aléatoires.
- L'issue dépend de l' état à l'instant t et pas de $t-1$.
- Loi des événements rares.

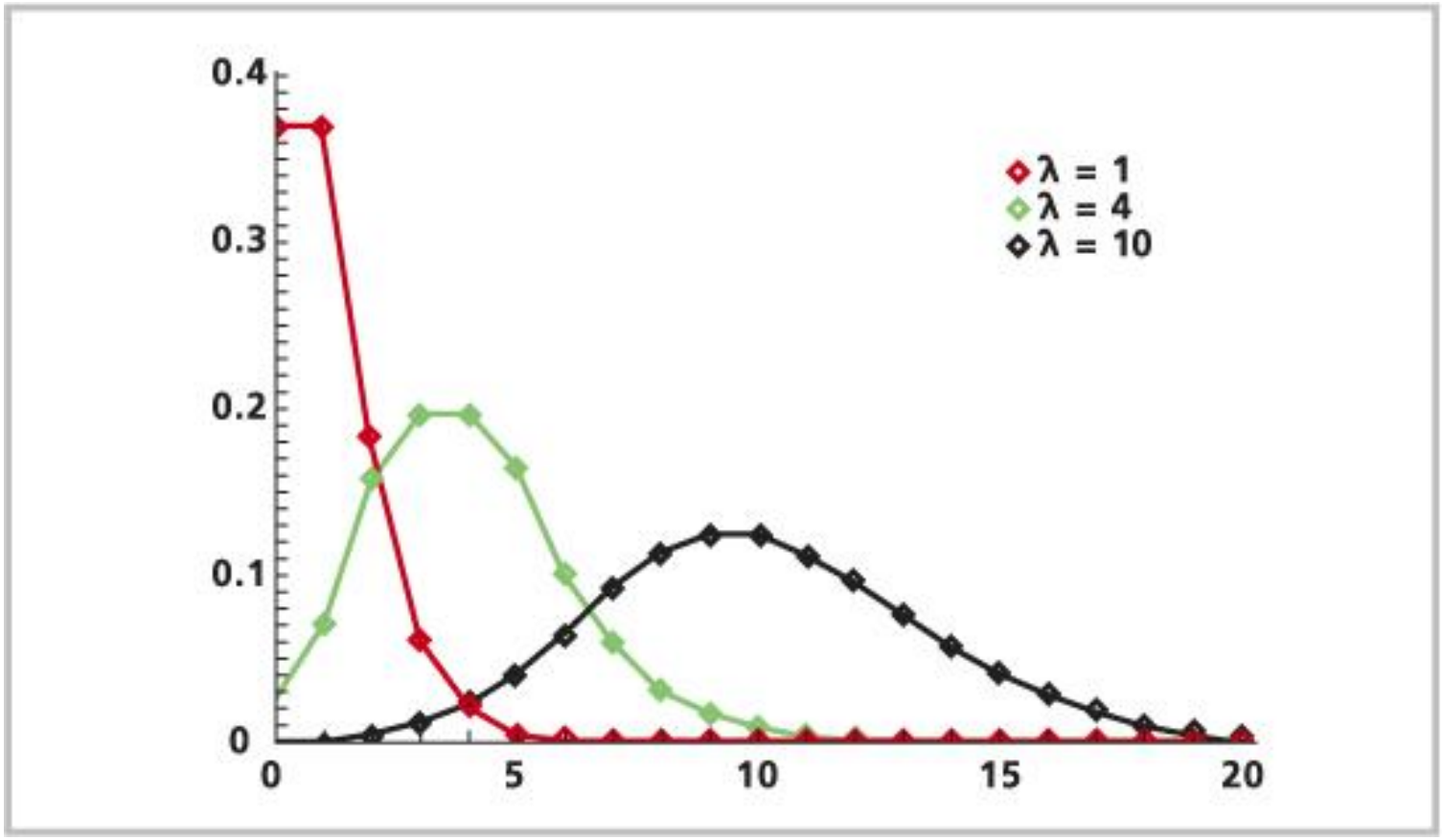
Moments d'une loi de Poisson

X variable discrète suit une loi de Poisson de paramètre $\lambda > 0$
$E(X) = V(X) = \lambda$

Commande R

- rpois(N,λ)

Loi de Poisson



Remarques

- Si $\lambda < 1$ la fonction de densité est dissymétrique.
- Plus λ est grand, plus la fonction de densité est symétrique.

Distribution de probabilités

Loi de Gauss ou loi normale

- Pour les Variables aléatoires continues.

Questions fondamentales :

- Les valeurs mesurées dans une expérience sont-elles celles d une grandeur aléatoire qui suit une loi normale ?
- La distribution de probabilités continue (fonction de densité) peut être utilisée pour de nombreux phénomènes que l on rencontre dans la nature.
- Une variable aléatoire X de moyenne μ et d'écart type σ suit une loi normale si sa densité de probabilité est : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\sigma)^2}{2\sigma^2}}$

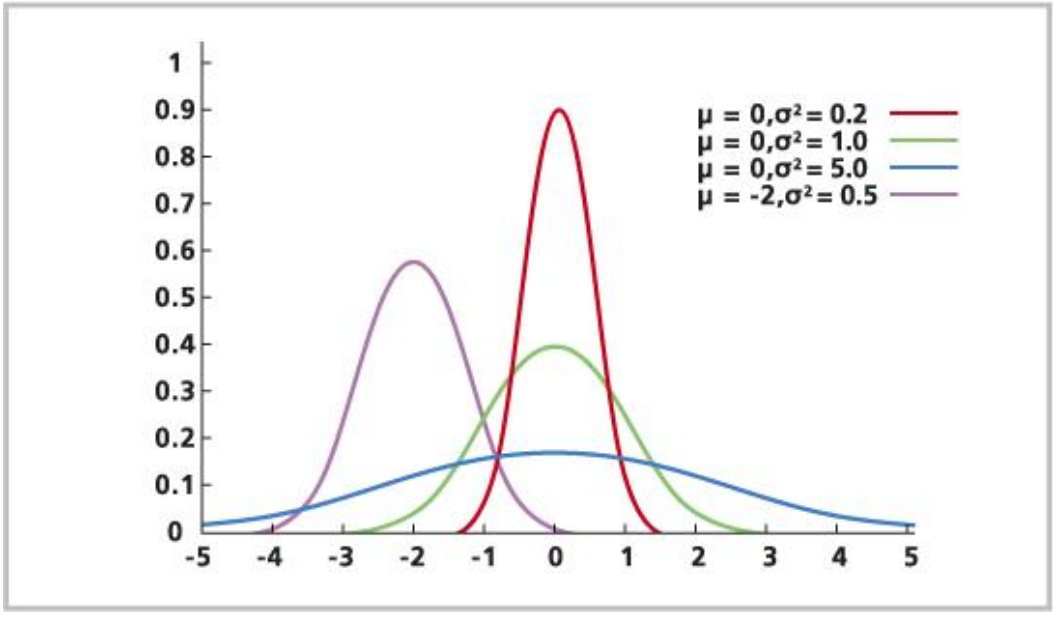
Commande R

rnorm(N, μ, σ)

Loi de Gauss ou loi normale

Moments de la loi de Gauss

- $E(X) = 0$
- $V(X) = 1$

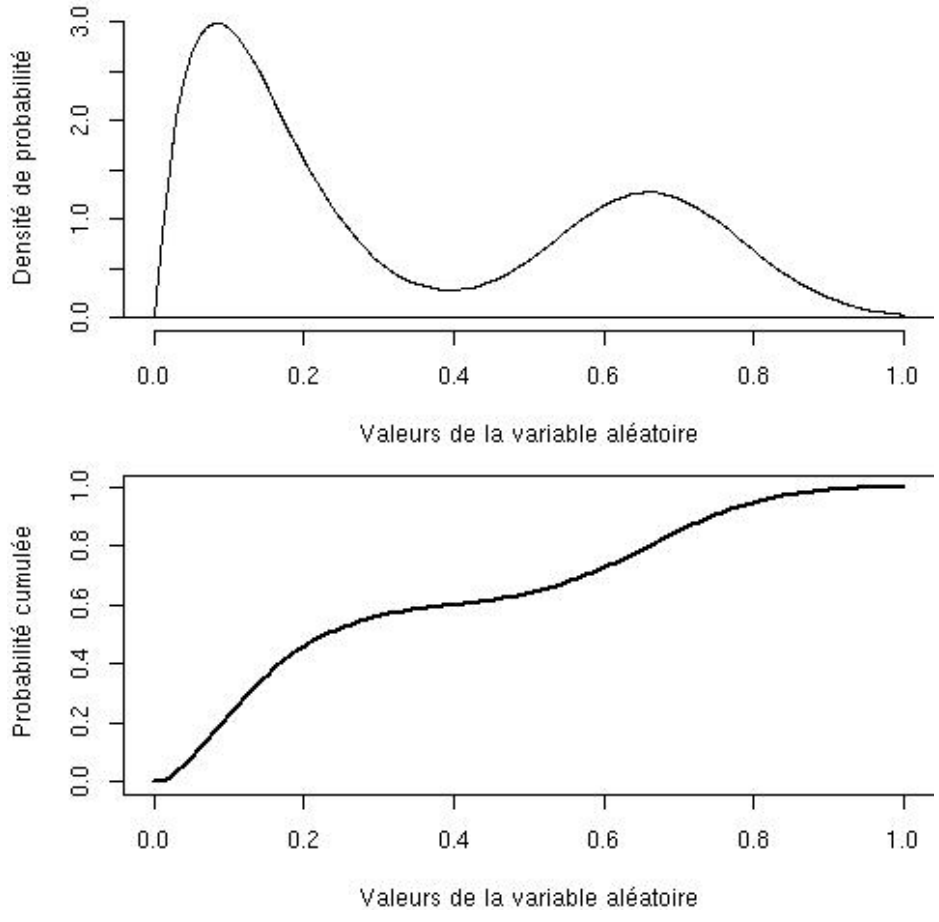


Remarques

Pour valider l'hypothèse d'une distribution normale on peut utiliser différents tests (Kolmogorov-Smirnov, Shapiro & Wilk ....)

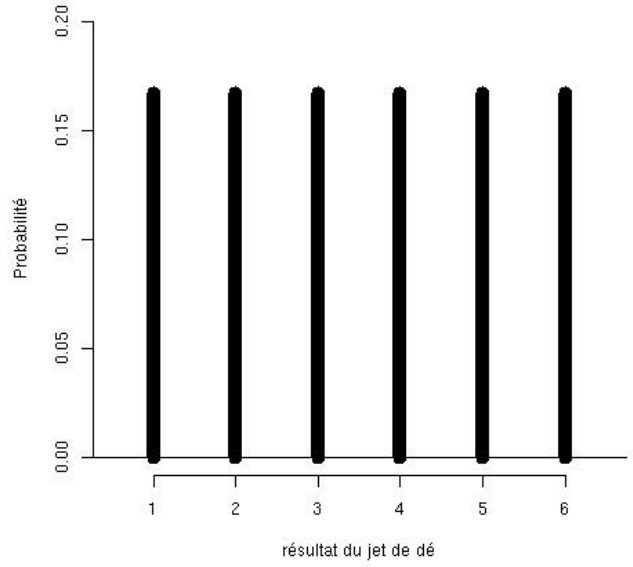
Distribution multimodale (variable continue)

Une variable aléatoire continue suit une loi multimodale lorsque la densité de probabilités montre un ou plusieurs maximum locaux.



Distribution uniforme (variable discrète ou continue)

Une distribution est uniforme lorsque la variable aléatoire X peut prendre n'importe quelle valeur entre une borne minimale et une borne maximale avec la même probabilité.
Par exemple, le jet de dé non truqué présente une distribution uniforme car la probabilité d'apparition de l'une des six faces est équiprobable.



Commande R

runif(*N*, *min*, *max*)