Variables aléatoires: Quelques lois usuelles

LMAFY1101

Anouar El Ghouch

LSBA, UCLouvain

Plan

LA LOI UNIFORME DISCRÈTE

On dit qu'une variable aléatoire X suit une loi uniforme discrète lorsqu'elle prend un nombre fini de valeurs, disons $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}$, et que ces valeurs sont équiprobables. Cela revient à dire que

$$P(X = x_i) = \frac{1}{n},\ i = 1, \ldots, n$$

On utilise cette loi pour modéliser des expériences aléatoires dont les résultats (événements élémentaires) sont équiprobables.

Il est facile de voir que

$$\mu_X = \frac{x_1 + \ldots + x_n}{n}$$
$$\sigma_X^2 = \frac{x_1^2 + \ldots + x_n^2}{n} - \mu_X^2$$

La loi Binomiale

La loi de Bernoulli

Une épreuve de Bernoulli est une expérience aléatoire avec seulement deux résultats possibles: S, pour succès, et E, pour échec. Soit p la la probabilité d'obtenir un S.

Fixons n, un entier positif, et considérons une séquence de n épreuves de Bernoulli indépendantes les unes des autres et identiques (ayant toutes la même probabilité de succès p). Posons

$$\boxed{X = \text{ le nombre de succès parmi } n}$$

Dans le cas particulier où $n = 1$, on dit que X suit une distribution de Bernoulli (ou simplement que X est une variable de Bernoulli) de paramètre p, et on écrit

$$X \sim Be(p)$$

Dans ce cas,

$$P(X = x) = p^x(1-p)^{1-x}, \text{ pour } x = 0, 1.$$

De la Bernoulli à la Binomiale

Dans le cas générale ($n \geqslant 1$), on dit que X suit une distribution Binomiale (ou simplement que X est une variable Binomiale) de paramètres (n, p), et on écrit

$$X \sim \mathrm{Bin}(n, p)$$

Il est facile de voir qu'une variable Binomiale n'est rien d'autre qu'une somme de variables de Bernoulli. En effet, si $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d. (indépendantes et identiquement distribuées) de Bernoulli $\mathrm{Be}(p)$, alors

$$\sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$$

La fonction de distribution de probabilité d'une v.a. $X \sim \mathrm{Bin}(n, p)$ est donnée par

$$P(X = x) = \underbrace{C_n^x}_{(1)} \underbrace{p^x}_{(2)} \underbrace{q^{n-x}}_{(3)}, \text{ pour } x = 0, 1, \ldots, n.$$

- (2) x fois succès, chacun avec une probabilité p
- (3) $n - x$ fois échec, chacun avec une probabilité $q = 1 - p$
- (1) $C_n^x = \dfrac{n!}{x!(n-x)!}$ est le nombre de résultats possibles avec x succès (et $(n-x)$ échecs) parmi n, càd. le nombre de combinaisons de x éléments choisis parmi n.

MOYENNE ET VARIANCE. $E(X) = np$ et $Var(X) = np(1-p)$

Excercice

Soit $X \sim \mathrm{Bin}(15, 0.3)$. Calculez $P(X = 5)$, $P(X \leqslant 5)$ et $P(X > 5)$.

Excercice

Soit $X \sim \mathrm{Bin}(15, 0.3)$. Calculez $P(X = 5)$, $P(X \leqslant 5)$ et $P(X > 5)$.

Solution. En appliquant la formule de la Binomiale on obtient

$$\begin{aligned} P(X = 5) &= C_{15}^{5} \times 0.3^{5} \times (1 - 0.3)^{10} = 0.2061304 \\ P(X \leqslant 5) &= \sum_{x=0}^{5} C_{15}^{x} 0.3^{x} 0.7^{(15-x)} = 0.7216214 \\ P(X > 5) &= 1 - P(X \leqslant 5) = 0.2783786 \end{aligned}$$

On peut faire ce type de calculs beaucoup plus facilement à l'aide de R.

La Binomiale dans R

Les fonctions dbinom, pbinom, rbinom

Soit $X \sim \mathrm{Bin}(n, p)$. Voici comment calculer des probabilités sur X à l'aide de R.

Prob.	Commande R
$P(X = x)$	`dbinom(x, size = n, prob = p)`
$P(X \leqslant x)$	`pbinom(x, size = n, prob = p)`
$P(X > x)$	`pbinom(x, size = n, prob = p, lower.tail = FALSE)`

Par exemple, voici $P(X = 5)$ pour $X \sim \mathrm{Bin}(15, 0.3)$,

```
> dbinom(5, size = 15, prob = 0.3)
```

```
[1] 0.206
```

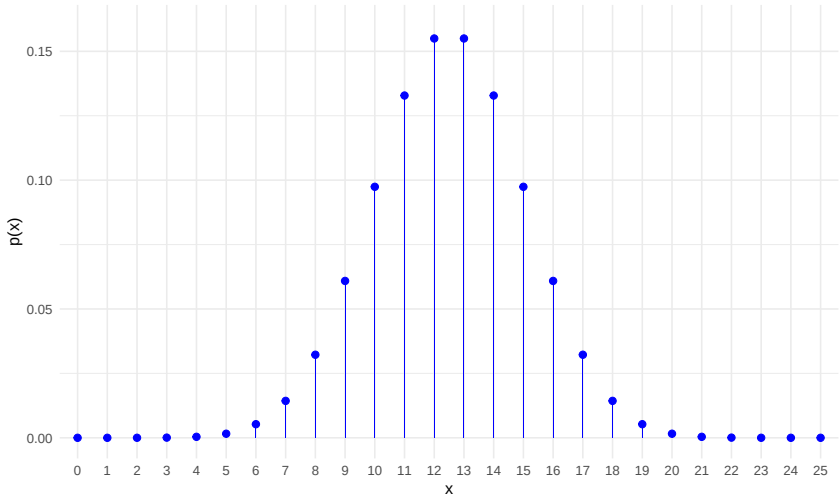
On peut aussi simuler des résiliations d'une Binomiale à l'aide de la fonction *rbinom*. Par exemple, voici comment générer 10 observations qui proviennent d'une $\mathrm{Bin}(15, 0.3)$.

```
> rbinom(10, size = 15, prob = 0.3)
```

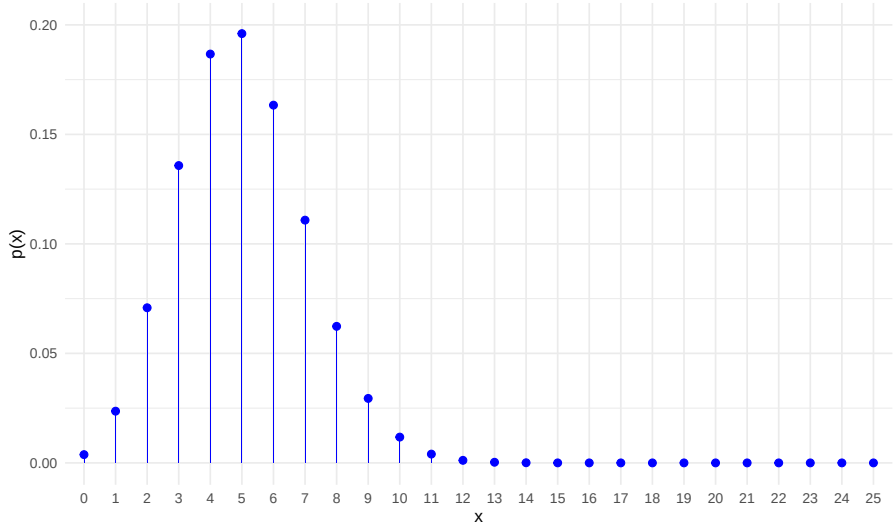
```
 [1] 6 3 4 4 3 4 4 3 5 4
```

La figure suivante montre la distribution $\mathrm{Bin}(25, 0.5)$

```
> x <- 0:25; y <- dbinom(x, size = 25, prob = 0.5)
> plot(y ~ x, type = "h"); points(x, y)
```



Voici un autre exemple avec $\mathtt{Bin}(25, 0.2)$



Loi de Poisson

La loi de Poisson permet de modéliser le nombre d'événements aléatoires survenant durant un intervalle (temporel ou spatiale) donné. Cela peut être un laps de temps limité (une heure, un jour, un mois, ...) ou un espace physique délimité (1 mètre, 2 litres, un certain volume, ...).

Par exemple:

- Le nombre de nouveaux cas de maladie par jours (dans une région donnée).
- Le nombre de particules émises par une source radioactive par seconde.
- Le nombre de pannes d'un réseau informatique par mois.
- Le nombre de bactéries présentes par ml dans une préparation biologique.
- Le nombre de fautes de frappe par page d'un livre.

Soit $X =$ nbr. d'occurrences d'un événement dans un intervalle donné.

Si nous formulons les hypothèses suivantes relativement à l'évenement qui nous intéresse:

- (Homogénéité) Le taux (moyenne) λ d'occurrence de l'événement reste fixe,
- (Indépendance) Les nombres de réalisations de l'événement au cours d'intervalles disjoints sont des v.a. indépendantes,
- La probabilité que l'événement se réalise plus d'une fois au cours d'un intervalle tend "rapidement" vers zéro au fur et à mesure que cet intervalle rétrécit.

alors on dit que X suit une distribution de Poisson de paramètre λ et on écrit $\boxed{X \sim \text{Pois}(\lambda)}$.
Dans ce cas,

$$\boxed{P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots}$$

Moyenne et variance. $E(X) = \lambda \quad \text{et} \quad Var(X) = \lambda$

Exemple

Les naissances à un hôpital ont lieu au hasard à un taux moyen de 1.8 naissance par heure. Quelle est la probabilité d'observer (a) 4 naissances en une heure, (b) plus que deux naissances en une heure et (c) 4 naissances en deux heures ?

Exemple

Les naissances à un hôpital ont lieu au hasard à un taux moyen de 1.8 naissance par heure. Quelle est la probabilité d'observer (a) 4 naissances en une heure, (b) plus que deux naissances en une heure et (c) 4 naissances en deux heures ?

Soit $X = \text{nbr. naissances durant 1h} \sim \text{Pois}(1.8)$
(a)

$$\text{P}(X = 4) = e^{-1.8}\frac{1.8^4}{4!} = 0.072$$

(b)

$$\text{P}(X > 2) = 1 - \text{P}(X \leqslant 2) = 1 - \text{P}(X = 0) - \text{P}(X = 1) - \text{P}(X = 2) = 0.27$$

(c) Soit $Y = \text{nbr. naissances durant 2h} \sim \text{Pois}(3.6)$

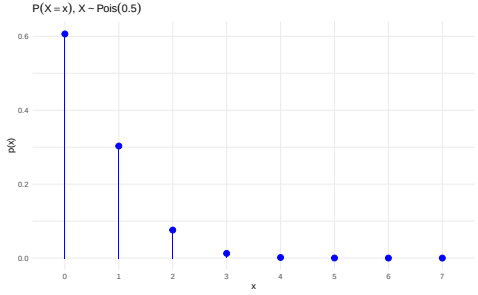
$$\text{P}(Y = 4) = e^{-3.6}\frac{3.6^4}{4!} = 0.19$$

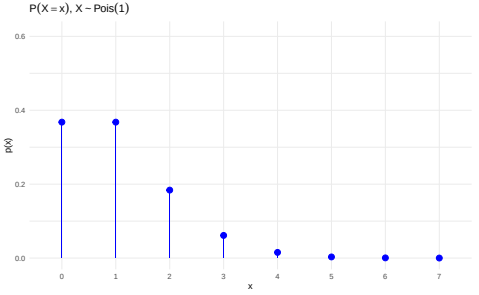
La distribution de Poisson dans R

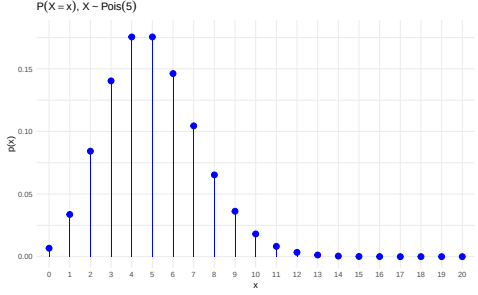
*Les fonctions d*pois*, p*pois*, r*pois

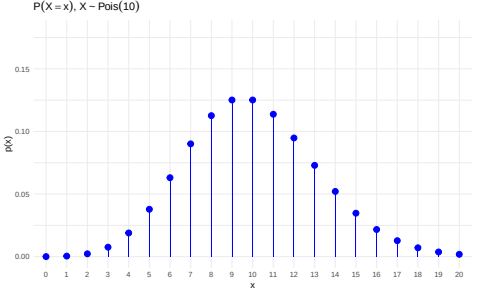
Suivant le même principe que pour la Binomiale, on peut utiliser R pour calculer les probabilités d'une variable Poisson. Pour cela il y a les fonctions *dpois*, *ppois* et *rpois*. Voici quelques exemples, avec une $X \sim \text{Pois}(1.8)$.

```
> # P(X = 4)
> dpois(4, lambda = 1.8)

[1] 0.0723


> # P(X > 2)
> ppois(2, lambda = 1.8, lower.tail = FALSE)

[1] 0.269


> # Générer 10 observations
> rpois(10, lambda = 1.8)

 [1] 2 2 1 3 2 0 1 3 1 1
```


P(X = x), X ~ Pois(0.5)
p(x)
x
0.0
0.2
0.4
0.6
0 1 2 3 4 5 6 7


P(X = x), X ~ Pois(1)
p(x)
x
0.0
0.2
0.4
0.6
0 1 2 3 4 5 6 7


P(X = x), X ~ Pois(5)
p(x)
x
0.00
0.05
0.10
0.15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20


P(X = x), X ~ Pois(10)
p(x)
x
0.00
0.05
0.10
0.15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

De la Binomiale à la Poisson

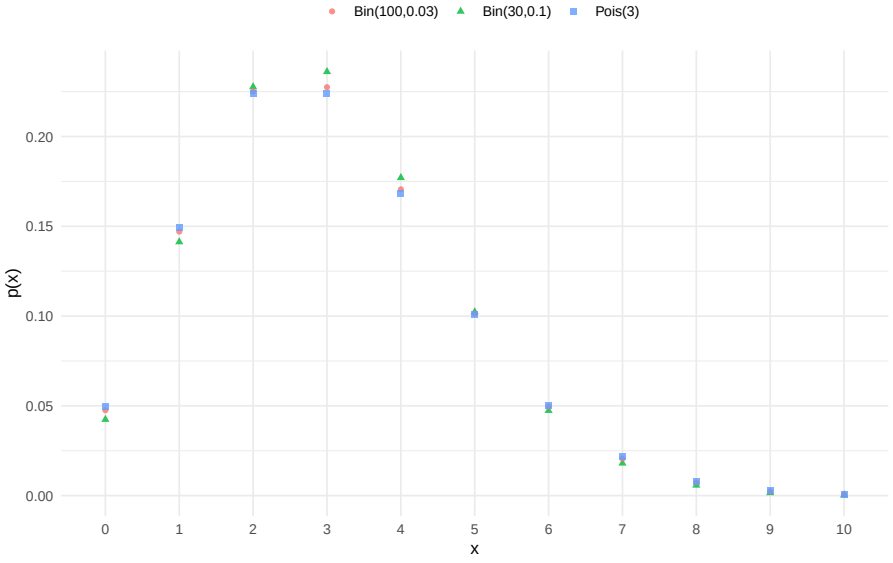
On peut montrer que la distribution de probabilité d'une Poisson est la limite d'une Binomiale. Plus précisément, on peut montrer que si $np_n \xrightarrow[n\to\infty]{} \lambda > 0$, alors

$$C_n^x p_n^x (1-p_n)^{n-x} \xrightarrow[n\to\infty]{} \frac{\lambda^x}{x!} e^{-\lambda}$$

C'est pour cette raison qu'on dit que la loi de Poisson est la loi des événements rares puisque il représente la limite d'une série de faits peu probables ($p_n \to 0$).

En pratique, lorsque n est "grand" et p est "petit" ($n > 50$ et $np < 5$) on peut remplacer la Binomiale par une Poisson

$$\mathrm{Bin}(n,p) \approx \mathrm{Pois}(np).$$


Bin(100,0.03)
Bin(30,0.1)
Pois(3)
p(x)
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
10
x

Binomiale ou Poisson ?

Ajuster une distribution aux données

Exemple 1

Voici les fréquences (Freq) de nombre d'enfants de sexe masculin ($n\mathrm{Males}$) dans 6115 familles. Chacune de ces familles est composée de 12 enfants.[1]

nMales	0	1	2	3	4	5	6	7	8	9	10	11	12
Freq	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Quel modèle (distribution) est approprié pour la variable $n\mathrm{Males}$?

(a) Binomiale ou (b) Poisson

[1] Étude réalisée en Saxe entre 1876 et 1885; Geissler, A. (*1889*)

Binomiale ou Poisson ?

Ajuster une distribution aux données

Exemple 1

Voici les fréquences (Freq) de nombre d'enfants de sexe masculin ($n\mathrm{Males}$) dans 6115 familles. Chacune de ces familles est composée de 12 enfants.[1]

nMales	0	1	2	3	4	5	6	7	8	9	10	11	12
Freq	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Quel modèle (distribution) est approprié pour la variable $n\mathrm{Males}$?

(a) Binomiale ou (b) Poisson

$n\mathrm{Males}$ ne peut prendre que les valeurs de 0 à 12 $\rightarrow$ la Binomiale semble plus plausible que la Poisson.

[1] Étude réalisée en Saxe entre 1876 et 1885; Geissler, A. (*1889*)

Question: le modèle Binomiale s'ajuste-t-il correctement aux données; càd. peut-on vraiment dire que $\mathtt{nMales} \sim \mathtt{Bin}(12, \mathrm{p})$, où p est la probabilité, pour une famille tirée au hasard, d'avoir un garçon ?

Question: le modèle Binomiale s'ajuste-t-il correctement aux données; càd. peut-on vraiment dire que $nMales \sim Bin(12, p)$, où p est la probabilité, pour une famille tirée au hasard, d'avoir un garçon ?

Pour répondre à cette question, on peut commencer par

- **Estimer** p par

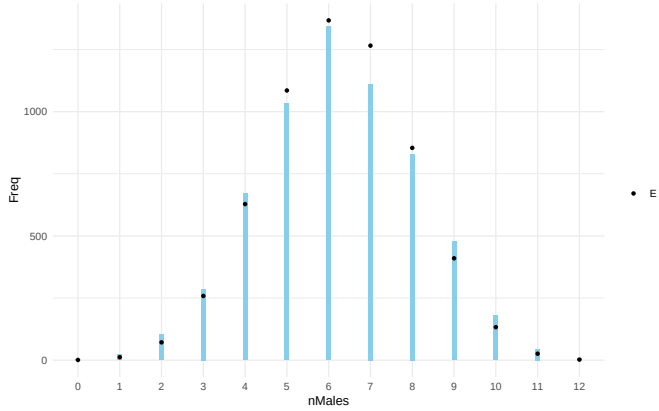
$$\hat{p} = \frac{\text{nombre total de garçons}}{\text{nombre total d'enfants}} = \frac{3 \times 0 + 24 \times 1 + \ldots 7 \times 12}{6115 \times 12} = 0.519215$$

- Puis comparer les probabilités théoriques $P(Bin(12, 0.5) = k)$, $k = 0, \ldots, 12$, aux proportions observées $Freq_k/6115$. Ce qui revient à comparer les effectifs observés $Freq_k$ aux **effectifs attendus** (E_k), sous l'hypothèse de la Binomial,

$$E_k = 6115 \times P(Bin(12, \hat{p}) = k),\ k = 0, \ldots, 12$$

Les résultats de ces calculs sont résumés dans le tableau et le graphique suivant.

nMales	Freq	E
0	3	0.93
1	24	12.09
2	104	71.80
3	286	258.48
4	670	628.06
5	1033	1085.21
6	1343	1367.28
7	1112	1265.63
8	829	854.25
9	478	410.01
10	181	132.84
11	45	26.08
12	7	2.35



Que peut-on conclure ?

Exemple 2

Voici les fréquences ($\mathtt{Freq}$) de nombre de naissances ($\mathtt{nBaby}$) par heure dans un hôpital. Ces données concernent une période de 24 heures.

nBaby	0	1	2	3	4
Freq	3	8	6	4	3

Quel modèle (distribution) est approprié pour $\mathtt{nBaby}$?

(a) Binomiale ou (b) Poisson

Exemple 2

Voici les fréquences ($\mathtt{Freq}$) de nombre de naissances ($\mathtt{nBaby}$) par heure dans un hôpital. Ces données concernent une période de 24 heures.

nBaby	0	1	2	3	4
Freq	3	8	6	4	3

Quel modèle (distribution) est approprié pour $\mathtt{nBaby}$?

(a) Binomiale ou (b) Poisson

$\mathtt{nBaby}$ peut prendre les valeurs $0, 1, 2, ... \rightarrow$ la distribution de Poisson peut être envisageable.

Question : le modèle Poisson s'ajuste-t-il correctement aux données : $\mathtt{nBaby} \sim \mathtt{Pois}(\lambda)$, pour un certain $\lambda > 0$?

Question : le modèle Poisson s'ajuste-t-il correctement aux données : $\mathtt{nBaby} \sim \mathrm{Pois}(\lambda)$, pour un certain $\lambda > 0$?

- Estimons λ:

$$\hat{\lambda} = \frac{\text{total des naissances}}{\text{temps total}} = \frac{8 \times 1 + 6 \times 2 + 4 \times 3 + 3 \times 4}{3 + 8 + 6 + 4 + 3} = 1.833$$

- Calculons les effectifs attendus (E) sous l'hypothèse de Poisson :

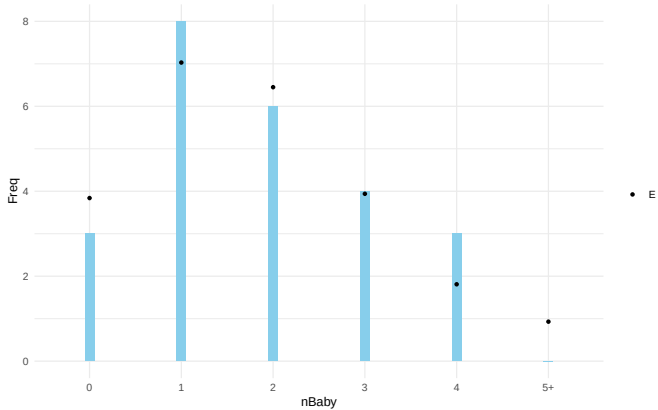
$$E_k = 24 \times P(\mathrm{Pois}(\hat{\lambda}) = k),\ k = 0, \ldots, 4$$

Puisque une Poisson $\in [0, \infty)$, il est logique d'inclure dans nos calculs la catégorie "5+", càd. celle qui correspond à $\mathtt{nBaby} > 4$.

$$E_{5+} = 24 \times P(\mathrm{Pois}(\hat{\lambda}) > 4) = 0.93$$

Comparons les effectifs observés (Freq) aux effectifs attendus.

nBaby	Freq	E
0	3	3.84
1	8	7.03
2	6	6.45
3	4	3.94
4	3	1.81
5+	0	0.93

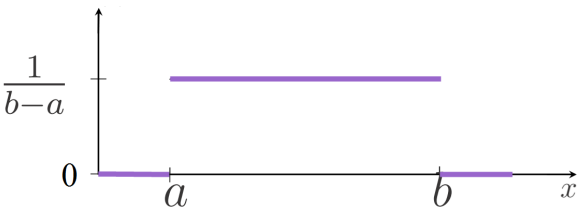


Que peut-on conclure ?

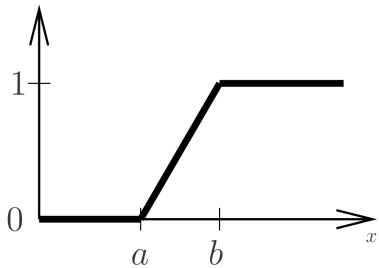
LA LOI UNIFORME CONTINUE

On dit qu'une variable aléatoire X suit une loi uniforme continue lorsqu'elle prend ses valeurs dans un intervalle fini, disons $[a, b]$, et tous les sous-intervalles de $[a, b]$ de même longueur sont équiprobables. Dans ce cas, on écrit $\boxed{X \sim \mathtt{Unif}(a, b)}$, ce qui revient à dire que la densité et la distribution cumulée de X sont données par

$$f(x) = \frac{1}{b-a} I(a \leqslant x \leqslant b)$$



$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x > b \end{cases}$$



MOYENNE ET VARIANCE. $E(X) = \frac{a+b}{2}$ et $Var(X) = \frac{(b-a)^2}{12}$

Les fonctions R *dunif*, *punif* permettent de calculer la densité et la distribution cumulée d'une uniforme. À titre d'exemple, voici comment calculer la probabilité qu'une $\mathtt{Unif}(5, 10)$ prenne une valeur inférieure à 7.5.

```
> punif(7.5, min = 5, max = 10)  # par défaut min = 0, et max = 1
```

```
[1] 0.5
```

Et voici un échantillon de 8 valeurs d'une $\mathtt{Unif}(5, 10)$

```
> runif(8, min = 5, max = 10)
```

```
[1] 6.33 6.86 7.86 9.54 6.01 9.49 9.72 8.30
```

Une autre fonction R utile est *qunif* qui n'est rien d'autre que la fonction quantile de l'uniforme, càd. la fonction $p \to a + p(b - a)$. Par exemple le quantile 0.5 d'une $\mathtt{Unif}(5, 10)$ est donné par

```
> qunif(0.5, min = 5, max = 10)
```

```
[1] 7.5
```

La loi normale

Il s'agit sans doute de la loi la plus importante, de par son ubiquité (à cause du théorème central limite).

X suit une loi normale (ou gaussienne), si elle a densité

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

$\pi \approx 3.14159$, $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$ sont les deux paramètres de cette distribution.
À partir de cette définition, on peut montrer que

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

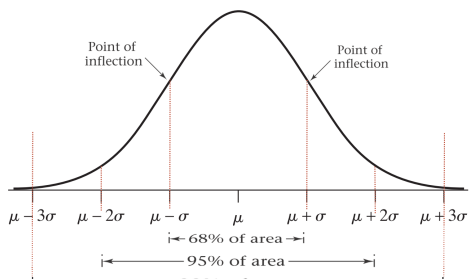
On écrit $\boxed{X \sim N(\mu, \sigma^2)}$ pour dire que X est une variable (de distribution) Normale de moyenne μ et de variance σ^2.

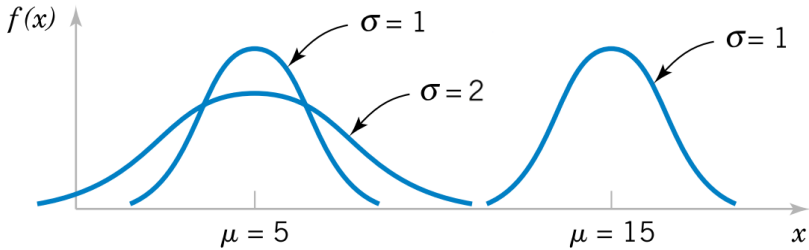
Cas particulier: La Normale standard, aussi appelée normale centrée réduite, correspond au cas où $\mu = 0$ et $\sigma = 1$. Donc une v.a. Z est normale standard, $Z \sim N(0, 1)$, si sa densité est

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad z \in \mathbb{R}$$

Caractéristiques d'une $N(\mu, \sigma^2)$

- densité en forme de cloche et s'étend de $-\infty$ à $+\infty$
- **symétrique et concentrée autour de sa moyenne**.
- de chaque côté, la densité décroît de façon exponentielle et tend vers 0





- μ contrôle la position de la courbe normale
- σ contrôle l'étalement de la courbe normale (la courbe s'aplatit plus rapidement lorsque $\sigma \nearrow$)
- Ainsi, un grand σ implique qu'une valeur de X loin de μ peut bien être observée, alors que cette valeur est assez peu probable quand σ est petit.

Lien entre $N(\mu, \sigma^2)$ et $N(0, 1)$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors, quels que soient les chiffres a et b, $a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$. Symboliquement, on peut écrire :

$$\boxed{a + b\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(a + b\mu, b^2\sigma^2)}$$

Conséquence

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

Cette opération est appelée **standardisation**. Son but est de récupérer une variable normale standard à partir d'une normale quelconque.

Application. Pour un $X \sim N(\mu, \sigma^2)$ et $Z \sim N(0, 1)$, on a

$$P(X \leqslant x) = P\left(\frac{X - \mu}{\sigma} \leqslant \frac{x - \mu}{\sigma}\right) = P(Z \leqslant (x - \mu)/\sigma)$$

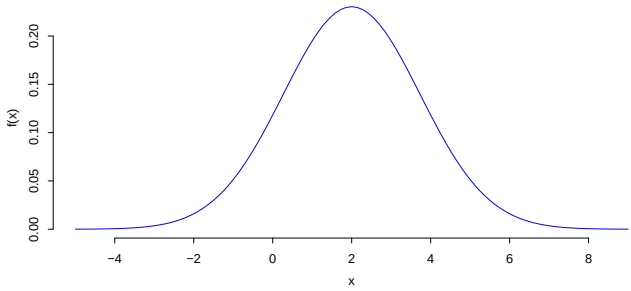
$$x_p = \mu + \sigma z_p,$$

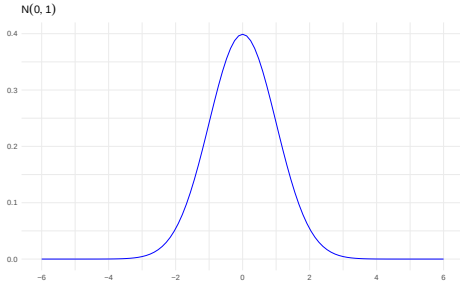
où x_p est le quantile d'ordre p d'une $\mathcal{N}(\mu, \sigma^2)$ et z_p est celui d'une $\mathcal{N}(0, 1)$.

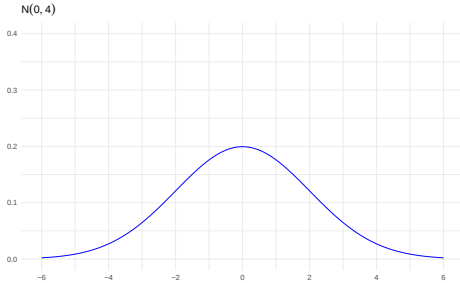
La distribution Normale dans R

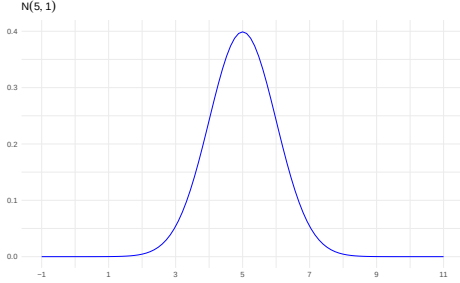
Les fonctions dnorm, pnorm, qnorm et rnorm

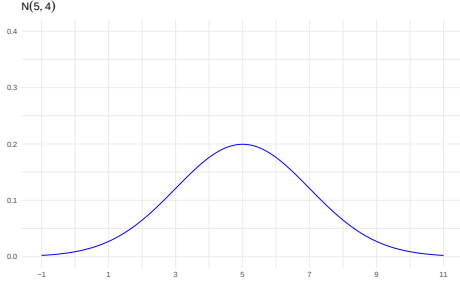
```
> # la fonction de densité d'une N(2,3) évaluée en 1
> dnorm(1, mean = 2, sd = sqrt(3))  # par défaut mean = 0, sd = 1

[1] 0.195

> # le graphe de la fonction de densité d'une N(2,3)
> curve(dnorm(x, mean = 2, sd = sqrt(3)), from = -5, to = 9, col = "blue", ylab = "f(x)")
```




N(0, 1)
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6

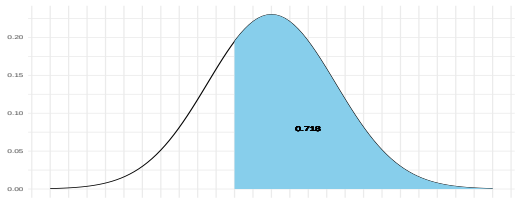

N(0, 4)
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6


N(5, 1)
0.4
0.3
0.2
0.1
0.0
−1
1
3
5
7
9
11


N(5, 4)
0.4
0.3
0.2
0.1
0.0
−1
1
3
5
7
9
11

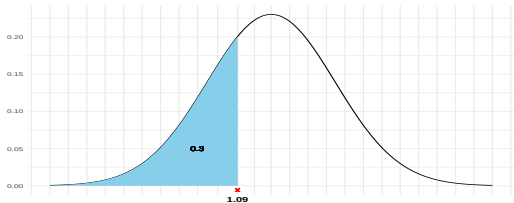
```
> # P(X > 1), X ~ N(2,3)
> pnorm(1, mean = 2, sd = sqrt(3), lower.tail = F)
```

```
[1] 0.718
```




```
> # quantile d'ordre 0.3 d'une N(2,3)
> qnorm(0.3, mean = 2, sd = sqrt(3))
```

```
[1] 1.09
```



Et voici comment simuler, par exemple, 10 observations d'une $\mathrm{N}(2,3)$

```
> rnorm(10, mean = 2, sd = sqrt(3))
```

```
 [1]  2.571  0.579  2.844  3.279  2.997  1.471  4.618  2.675  0.924 -1.836
```

Exercice 1

À l'aide de R calculez $P(Z < 1.37)$. En déduire (a) $P(Z > 1.37)$, (b) $P(Z > -1.37)$ et (c) $P(Z < -1.37)$

Exercice 1

À l'aide de R calculez $P(Z < 1.37)$. En déduire (a) $P(Z > 1.37)$, (b) $P(Z > -1.37)$ et (c) $P(Z < -1.37)$

```
> pnorm(1.37)

[1] 0.915
```

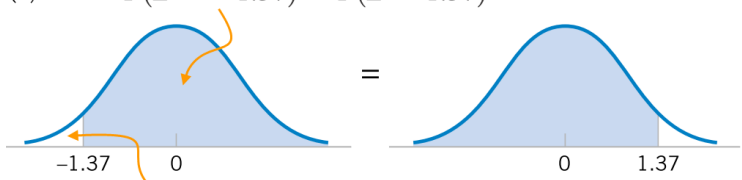
Exercice 1

À l'aide de R calculez $P(Z < 1.37)$. En déduire (a) $P(Z > 1.37)$, (b) $P(Z > -1.37)$ et (c) $P(Z < -1.37)$

```
> pnorm(1.37)

[1] 0.915
```

(a) $P(Z > 1.37) = 1 - P(Z < 1.37) = 0.085$

(b) $P(Z > -1.37) = P(Z < 1.37) = 0.915$



(c) $P(Z < -1.37) = P(Z > 1.37) = 0.085$

Exercice 2

À l'aide de R calculez $z_{0.85}$ le quantile 0.85 d'une $\mathcal{N}(0,1)$. En déduire $z_{0.15}$

Exercice 2

À l'aide de R calculez $z_{0.85}$ le quantile 0.85 d'une $\mathcal{N}(0,1)$. En déduire $z_{0.15}$

```
> qnorm(0.85)
```

```
[1] 1.04
```

Exercice 2

À l'aide de R calculez $z_{0.85}$ le quantile 0.85 d'une $\mathcal{N}(0,1)$. En déduire $z_{0.15}$

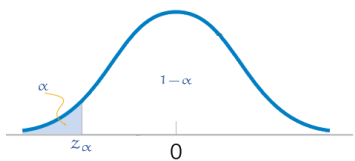
```
> qnorm(0.85)

[1] 1.04
```

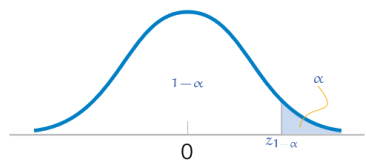
$z_{0.15}$ est $-z_{0.85} = -1.04$. En effet,

```
> qnorm(0.15)

[1] -1.04
```

De façon générale

$$z_{1-\alpha} = -z_\alpha$$





Exercice 3

Soit $X \sim N(10, 4)$. Utilisez la fonction pnorm sans les paramètres $mean$ et sq pour (a) calculer $P(8 \leqslant X \leqslant 13)$, (b) trouvez x tel que $P(X \leqslant x) = 0.2$.

Exercice 3

Soit $X \sim N(10, 4)$. Utilisez la fonction pnorm sans les paramètres $mean$ et sq pour (a) calculer $P(8 \leqslant X \leqslant 13)$, (b) trouvez x tel que $P(X \leqslant x) = 0.2$.

(a)

```
> pnorm((13 - 10)/2) - pnorm((8 - 10)/2)
```

```
[1] 0.775
```

Ce qui est le même que `pnorm(13, mean = 10, sd = 2) - pnorm(8, mean = 10, sd = 2)` .

(b)

```
> 10 + 2 * qnorm(0.2)
```

```
[1] 8.32
```

Ce qui est le même que `qnorm(0.2, mean = 10, sd = 2)` .

Exercice 4 (Règle des trois sigmas)

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. (a) Montrer que

$$P(\mu - 1.96 \times \sigma \leqslant X \leqslant \mu + 1.96 \times \sigma) = 0.95$$

(b) Trouvez l'intervalle $[a, b]$ le plus petit possible tel que

$$P(X \in [a, b]) = 0.99$$

Exercice 4 (Règle des trois sigmas)

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. (a) Montrer que

$$P(\mu - 1.96 \times \sigma \leqslant X \leqslant \mu + 1.96 \times \sigma) = 0.95$$

(b) Trouvez l'intervalle $[a, b]$ le plus petit possible tel que

$$P(X \in [a, b]) = 0.99$$

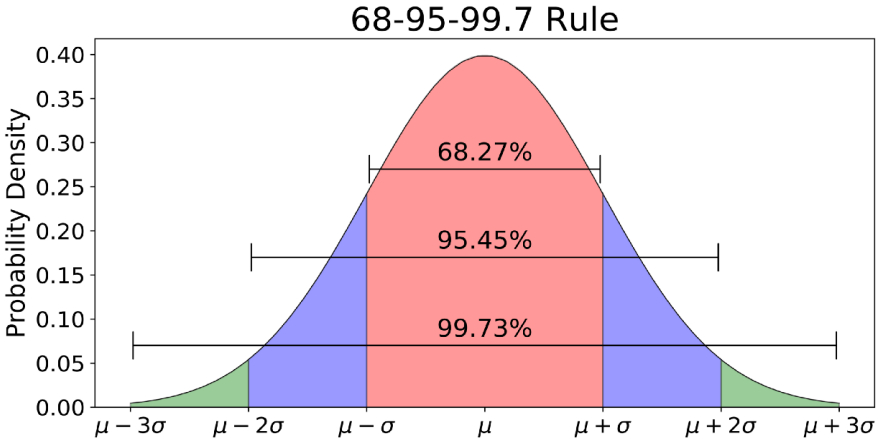
Solution.
(a)

$$P(\mu - 1.96 \times \sigma \leqslant X \leqslant \mu + 1.96 \times \sigma) = P(-1.96 \leqslant Z \leqslant 1.96) = 2P(Z \leqslant 1.96) - 1 = 0.95.$$

(b)

$$[a, b] = [\mu - 2.58 \times \sigma, \mu + 2.58 \times \sigma]$$

C'est ce type de calculs qui nous permet de prouver la règle dite de $68-95-99.7$ (ou règle des trois sigmas) qui indique que pour une loi normale, presque toutes les valeurs se situent dans un intervalle centré autour de la moyenne et dont les bornes se situent à 3 écarts-types de part et d'autre.



Soit X_1 et X_2 deux v.a. **indépendantes**, si $X_1 \sim N(\mu_1, \sigma_1^2)$ et $X_2 \sim N(\mu_2, \sigma_2^2)$, alors $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_1^2)$.

Symboliquement, on peut écrire : $N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

De façon générale, en cas d'indépendance,

$$\sum_{i=1}^{n} N(\mu_i, \sigma_i^2) = N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

En particulier, $\sum_{i=1}^{n} N(\mu, \sigma^2) = N\left(n\mu, n\sigma^2\right)$.

Excrcice. Soit $X_1 \sim \mathcal{N}(0, 1)$ et $X_2 \sim \mathcal{N}(1, 9)$. On suppose que X_1 et X_2 sont indépendantes. Calculez à la main $P(3X_1 + X_2 \leqslant 1)$.

Distribution d'échantillonnage

Pour bien comprendre l'idée de base considérant l'expérience suivante: on lance un dé trois fois et on note la moyenne des trois chiffres obtenus.

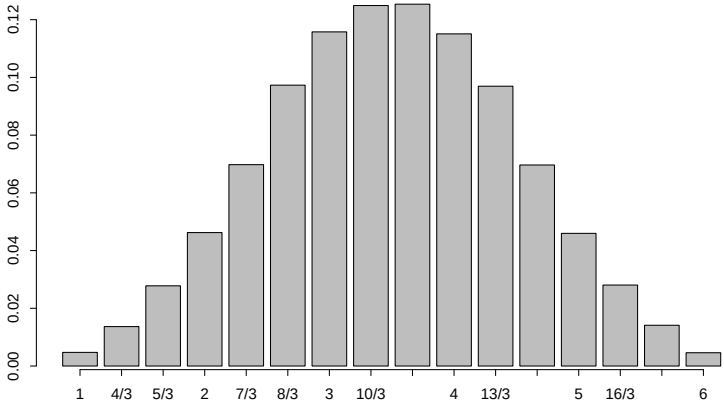
Nous pouvons modéliser cette expérience de la façon suivante. Soit X_i = la face obtenue pour le i—ème dé, $i = 1, 2, 3$. Ces X_i sont des v.a. indépendantes et identiquement distribuées (i.i.d.). Soit

$$\overline{X}_3 = \frac{X_1 + X_2 + X_3}{3}$$

$\overline{X}_3$ est une v.a. qui se modifie en fonction des valeurs prises par X_1 ,X_2 et X_3.

Dans ce cas particulier nous pouvons calculer la distribution de $\overline{X}_3$ à la main ou à l'aide de R.

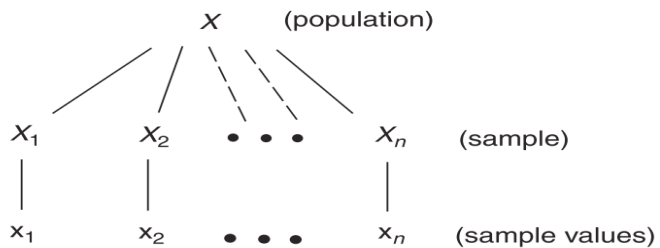
```
> exp <- replicate(10^6, {
    smp <- sample(1:6, 3, replace = TRUE)
    mean(smp)
  })
> table(exp) |> proportions() |> barplot()
```



De façon générale, soit X une v.a. de moyenne μ et de variance σ^2. Cette v.a. est le résultat d'une expérience aléatoire.

Si cette dernière est répétée n fois de manière identique et indépendante, on obtient n observation/valeurs : $x_1, x_2, \ldots, x_n$.

Chaque x_i n'est rien d'autre qu'une réalisation de la v.a. X_i : "résultat de la i—ème expérience".



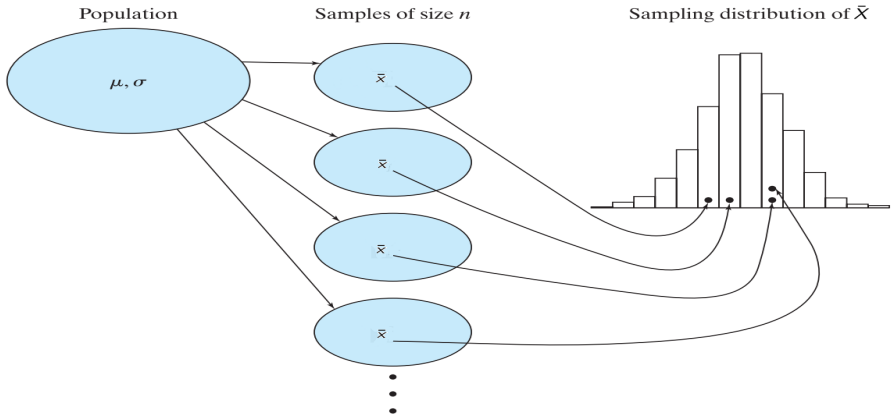
Population, sample, and sample values

Les $X_1, \ldots, X_n$ est une suite de v.a. **indépendantes et identiquement distribuées** (**i.i.d.**). Chaque X_i a la même distribution que X.

Nous appellerons $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ la moyenne de l'échantillon aléatoire $\{X_1, \ldots, X_n\}$.

$\overline{X}_n$ **est une variable aléatoire** puisque *il change de valeurs en fonction des mesures récoltées lors de l'échantillonnage*, càd en fonction des valeurs prises par $X_1, \ldots, X_n$.

Si on considère un échantillon particulier $\{x_1, \ldots, x_n\}$ et que l'on calcul sa moyenne $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, on obtiendrait une réalisation, parmi d'autres, de $\overline{X}_n$. Si l'on réalise une autre série de n nouvelles mesures, nous obtiendrons une nouvelle moyenne qui représenterait une autre réalisation de $\overline{X}_n$, et ainsi de suite.



Comme toute variable aléatoire $\overline{X}_n$ dispose d'une moyenne, d'une variance et d'une distribution dite **distribution d'échantillonnage** (en anglais, sampling distribution).

Il est facile de voir que, quelque soit la distribution de X,

$$E(\overline{X}_n) = \mu \quad \text{et} \quad Var(\overline{X}_n) = \sigma^2/n$$

La distribution exacte de $\overline{X}_n$ est inconnue sauf dans des cas bien particuliers.

Cas d'une population Normale: Si $X \sim N(\mu, \sigma^2)$, càd si $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, alors

$$\boxed{\overline{X}_n \sim N(\mu, \sigma^2/n)}$$

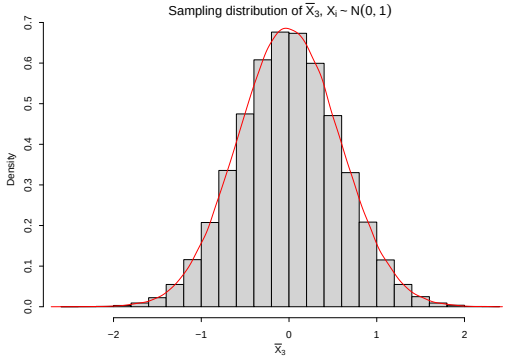
Si les $X_i \not\sim N$, alors on ne sait pas dire si $\overline{X}_n$ est normale ou pas.

Illustration

Simuler la distribution de $\overline{X}_3$ dans le cas d'une

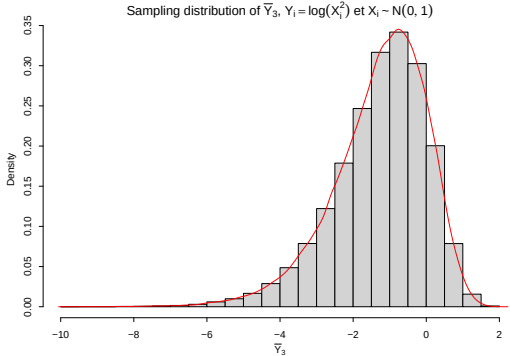
Population Normale

```
> barX3 <- replicate(100000, rnorm(3) |> mean())
> barX3 |> hist(freq = FALSE)
> barX3 |> density() |> lines(col = "red")
```



Population pas Normale

```
> barY3 <- replicate(100000, log(rnorm(3)^2) |> mean())
> barY3 |> hist(freq = FALSE)
> barY3 |> density() |>  lines(col = "red")
```



Théorème central limite (CLT)

CLT est l'un des résultats les plus importants en statistique. Il justifie l'importance accordée à la loi normale.

Intuitivement, le théorème central limite affirme que toute somme de variables aléatoires indépendantes tend, dans certains cas, vers une normale.

Si $X_1, \ldots, X_n$ est une suite de v.a. **indépendantes et identiquement distribuées** (**i.i.d.**) suivant la même loi, alors quand n est assez grand,

$$\sum_{i=1}^{n} X_i \text{ suit approximativement une loi } \mathcal{N}(n\mu, n\sigma^2),$$

où $\mu = E(X_i)$ et $\sigma^2 = Var(X_i)$.

On peut exprimer le CLT en écrivant que, pour de grandes valeurs de n,

$$\sum_{i=1}^{n} X_i \sim_a \mathcal{N}(n\mu, n\sigma^2), \text{ ou, de façon équivalente,}$$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim_a \mathcal{N}(\mu, \sigma^2/n)$$

Illustration 1

```
> barY3 <- replicate(100000, log(rnorm(100)^2) |> mean())
> barY3 |> hist(freq = FALSE)
> barY3 |> density() |>  lines(col = "red")
```

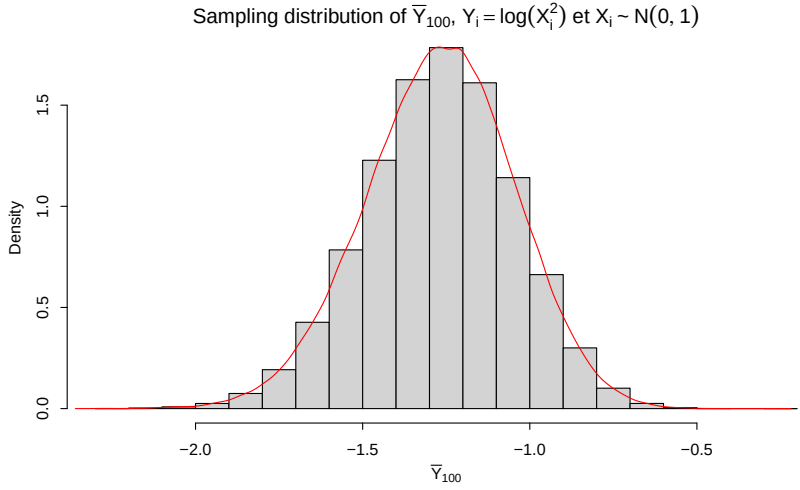
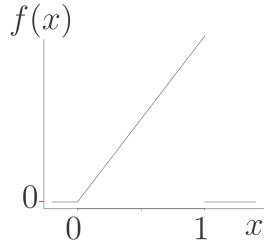


Illustration 2

Soit X_i la v.a. de densité $f(x) = 2x$ pour $0 < x < 1$. On peut facilement vérifier que $\mathbb{E}(X_i) = 2/3$ et $\mathbb{V}\mathrm{ar}(X_i) = 1/18$.
Par CLT, $S_n = \sum_{i=1}^{n} X_i \sim_a \mathcal{N}(2n/3, n/18)$.



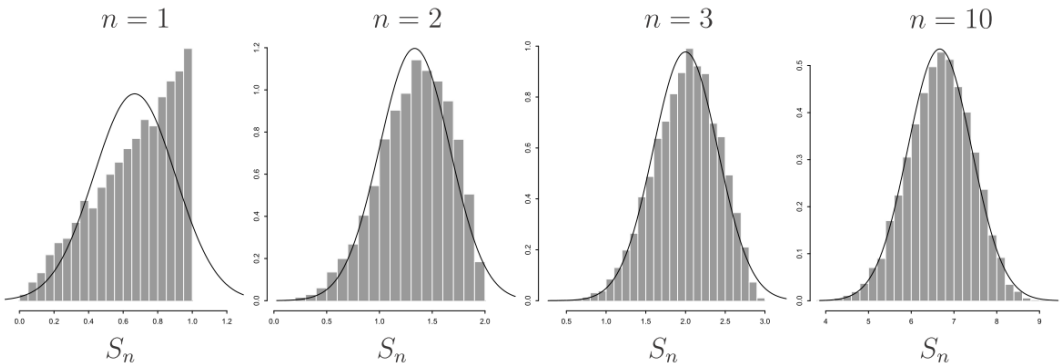
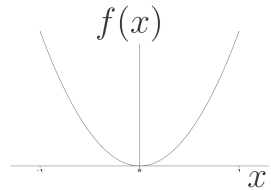
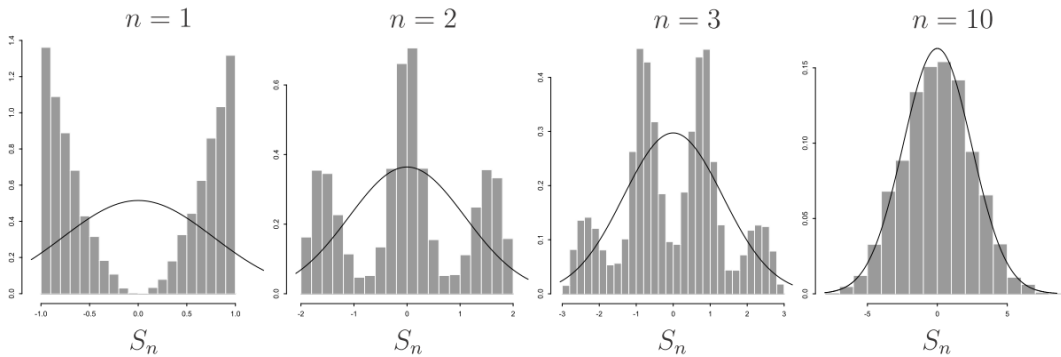


Illustration 3

Soit X_i la v.a. de densité $f(x) = \frac{3}{2}x^2$ pour $-1 < x < 1$. On peut facilement vérifier que $\mathbb{E}(X_i) = 0$ et $\mathbb{V}ar(X_i) = 3/5$.
Par CLT, $S_n = \sum_{i=1}^{n} X_i \sim_a \mathcal{N}(0, 3n/5)$.





Approximation d'une binomiale par une normale

Soit $X \sim Bin(n, p)$, on a vu que X peut s'écrire comme la somme de v.a. indépendantes de Bernoulli, c.à.d.

$$X = X_1 + X_2 + \ldots + X_n,$$

avec $X_i \sim Be(p)$, puisque $\mathbb{E}(X_i) = p$ et $\mathbb{V}ar(X_i) = pq$ $(q = 1 - p)$, alors par CLT

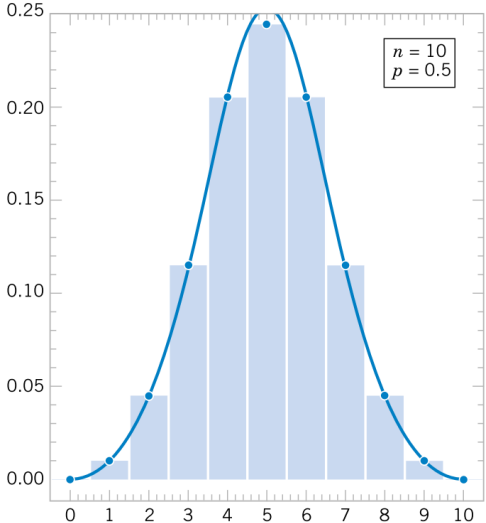
$$X \sim_a \mathcal{N}(np, npq)$$

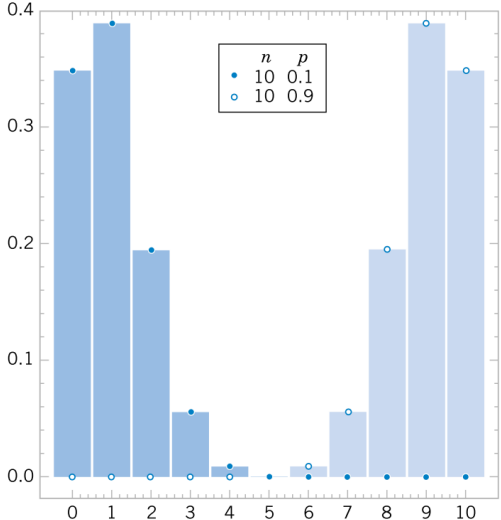
Cette approximation sera d'autant meilleure que p est proche de $1/2$ et un grand n.

Symboliquement, on peut écrire, pour n grand,

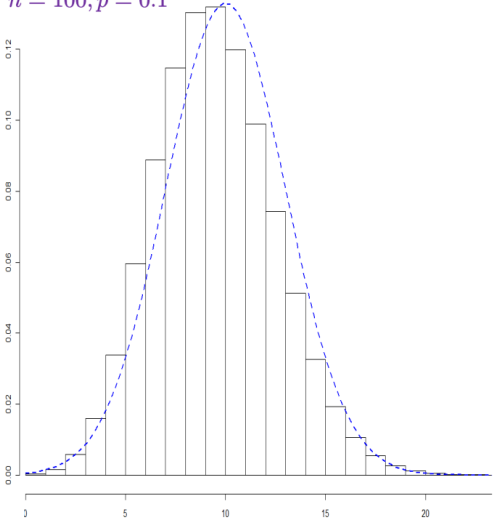
$$\boxed{Bin(n, p) \approx \mathcal{N}(np, npq)}$$

Illustration

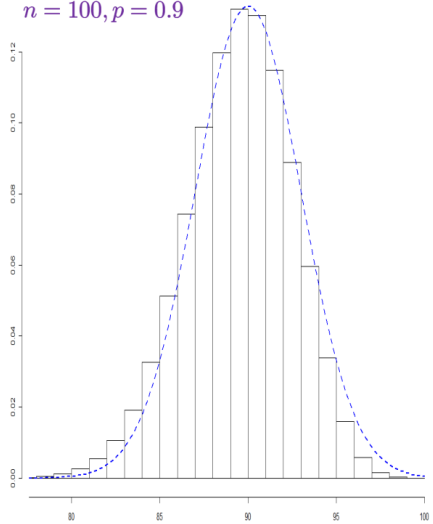




$n = 100, p = 0.1$



$n = 100, p = 0.9$

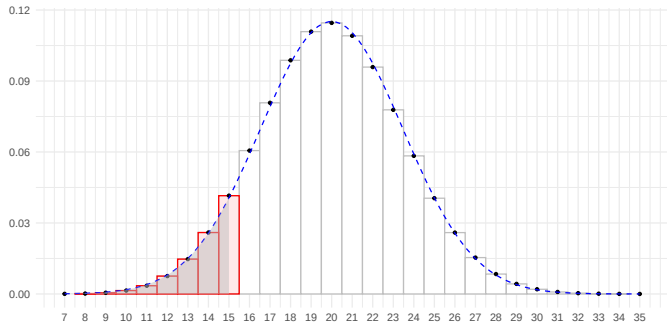


Exemple

Une expérience de Bernoulli est répétée 50 fois, chaque expérience ayant une probabilité de succès $p = 0.4$. Calculons la probabilité d'obtenir moins que 15 succès.

$\mathrm{P}(\mathrm{Bin}(50, 0.4) \leqslant 15) =$ `pbinom(15, 50, 0.4)` $=$ `0.096`

$\mathrm{P}(\mathrm{N}(50 \times 0.4, 50 \times 0.4 \times 0.6) \leqslant 15) =$ `pnorm(15, 20, sqrt(12))` $=$ `0.074`



Lorsque nous utilisons la loi normale (continue) pour approximer la binomiale (discrète), nous approximons en faite la surface (hauteur) des bâtons par la surface sous la courbe de la densité normale.

Souvent, on peut améliorer cette approximation en appliquant la correction dite "de continuité" qui consiste à remplacer une probabilité de type $\mathrm{P}(\mathrm{Bin} \leqslant \mathrm{x})$ par $\mathrm{P}(\mathrm{N} \leqslant \mathrm{x} + 0.5)$ et une probabilité de type $\mathrm{P}(\mathrm{Bin} > \mathrm{x})$ par $\mathrm{P}(\mathrm{N} > \mathrm{x} + 0.5)$, de telle sorte à intégrer dans la surface de la normale la partie manquante (voir figure ci-dessus).

Dans notre cas cela donne

```
pnorm(15 + 0.5, 20, sqrt(12)) = 0.097
```

CLT : quelle valeur de n est assez grande ?

La valeur de n à partir de laquelle le CLT aboutit à une bonne approximation dépend de la distribution des X_i. Si cette dernière est

- "très proche" d'une normale (symétrique atour de sa moyenne, avec un seul mode et sans valeurs extrêmes) alors un $n \geqslant 5$ est suffisant.
- "moyennement proche" d'une normale, alors il faut un $n \geqslant 30$.
- "loin d'une normale" (fortement asymétrique, par exemple), alors il faut un $n \geqslant 100$.

Conformité à la loi normale

- L'hypothèse de normalité (l'adéquation des données à la normale) est souvent requise dans de nombreuses techniques d'inférence statistique (tests, intervalles de confiance, etc).
- Si c'est le cas, alors il faut toujours vérifier la normalité avant d'effectuer l'analyse au risque de tirer des conclusions erronées.
- Il existe plusieurs outils pour vérifier la normalité. Ici, on se limitera à des techniques descriptives (graphiques) qui ne permettent pas de prouver la normalité en toute rigueur, mais de vérifier si la distribution des données est réellement incompatible avec la distribution normale (ex. asymétrie forte, distribution avec plusieurs modes, etc.).

Point de départ

On dispose de n observations $x_1, \ldots, x_n$ qui proviennent de X.

La problématique est la suivante :

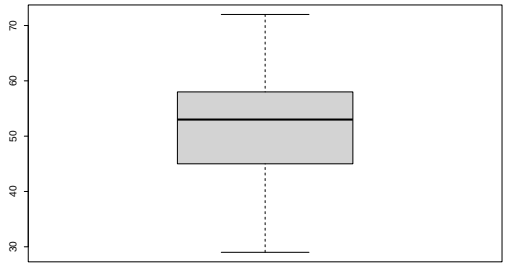
Est-ce que les observations dont nous disposons nous permettent d'affirmer que X suit (ou pas) une loi normale ?

Exemple. On réalise $n = 50$ expériences chimiques de manières identiques et indépendantes. Le temps de réaction (X) en secondes observé pour chacune des expériences est :

43	48	65	55	51	44	51	59	62	51
45	53	55	55	49	34	52	69	45	54
59	36	36	29	52	59	41	58	54	55
72	53	52	49	57	42	70	58	42	53
57	68	40	65	54	49	32	56	50	59

Box-plot

```
> x <- c(43, 48, 65, 55, 51, 51, 44, 51, 59, 62, 45, 53, 55, 55, 49, 34, 52, 69,
      45, 54, 59, 36, 36, 29, 52, 59, 41, 58, 54, 55, 72, 53, 52, 49, 57, 42, 70,
      58, 42, 53, 57, 68, 40, 65, 54, 49, 32, 56, 50, 59)
> boxplot(x)
```

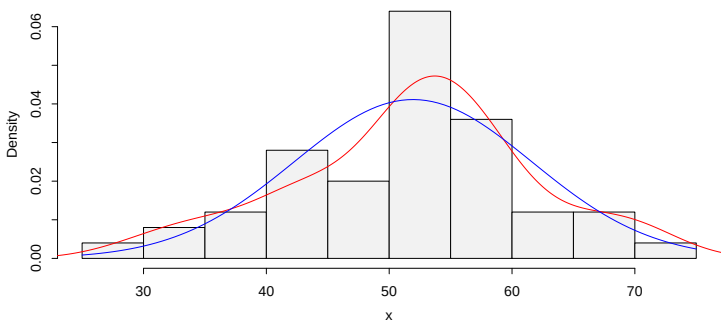


Le box-plot permet de visualiser rapidement la forme (la symétrie) de la distribution des données et la présence de valeurs atypiques.

Histogramme/densité

On trace l'histogramme et/ou la densité des données. Puis on superpose les valeurs de la densité normale théorique en estimant les paramètres inconnus de celle-ci, à savoir μ et σ^2, par la moyenne et la variance empiriques.

```
> hist(x, freq = FALSE, col = "grey95", main = "")
> lines(density(x), col = "red")
> curve(dnorm(x, mean = 51.94, sd = 9.7), add = T, col = "blue")
```



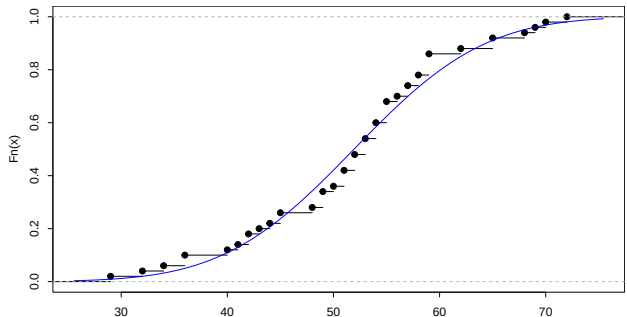
Fonction de répartition empirique

On trace la fonction de répartition empirique associée à notre échantillon $\{x_1, \ldots, x_n\}$

$$\hat{F}_n(x) = \frac{\text{nombre de } x_i \leqslant x}{n} = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leqslant x)$$

Puis on superpose le graphe de la fonction de répartition de la normale en estimant ses paramètres par la moyenne et la variance empiriques.

```
> plot(ecdf(x), main = "")
> curve(pnorm(x, mean = 51.94, sd = 9.7), add = T, col = "blue")
```

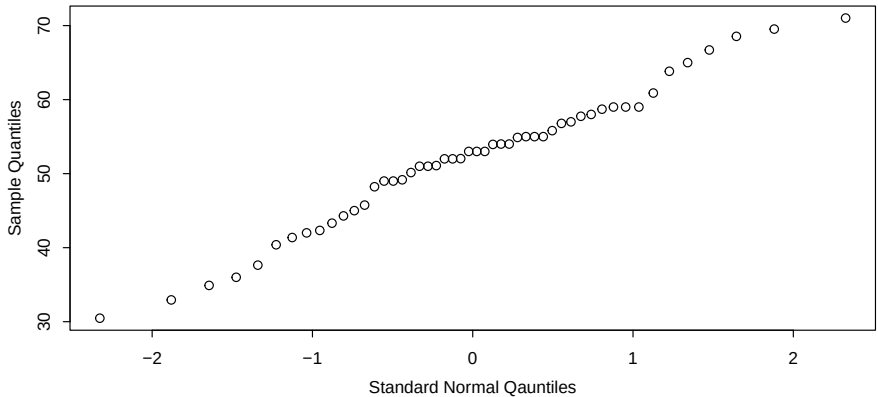


- Soit z_p le quantile d'ordre p d'une v.a. $Z \sim \mathcal{N}(0,1)$. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors le quantile d'ordre p de X vérifie

$$x_p = \mu + \sigma z_p$$

- Par conséquent, si on note par $\hat{x}_p$ le quantile empirique d'ordre p calculé à partir de l'échantillon $\{x_1, x_2, \ldots, x_n\}$, alors le nuage de points $\{(z_p, \hat{x}_p),\ 0 < p < 1\}$ doit dessiner plus ou moins une droite.

```
> zp <- qnorm(seq(0.01, 0.99, 0.02))
> xp <- quantile(x, seq(0.01, 0.99, 0.02))
> plot(x = zp, y = xp, xlab = "Standard Normal Qauntiles", ylab = "Sample Quantiles")
```



La fonction *qqnorm* permet de tracer facilement un qq-plot. R possède également la fonction *qqline*, qui ajoute une droite au graphique. Cette droite facilite beaucoup l'évaluation de la normalité des données. Plus tous les points sont proches de la ligne, plus la distribution de l'échantillon se rapproche de la distribution normale.

```
> qqnorm(x)
> qqline(x)
```

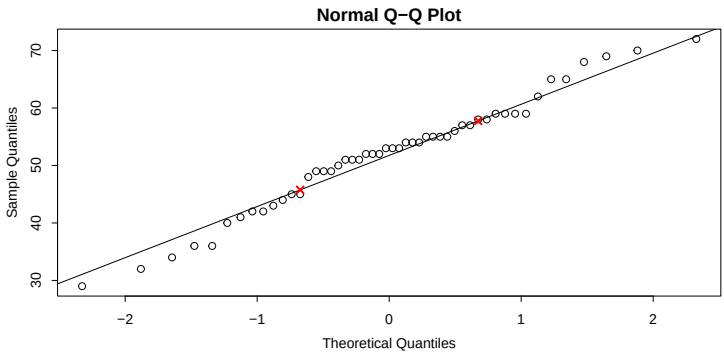
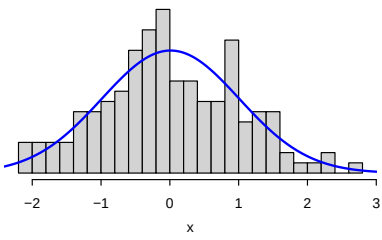
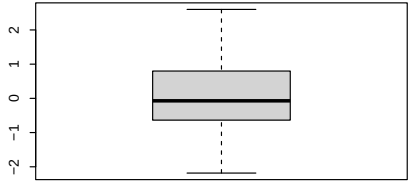
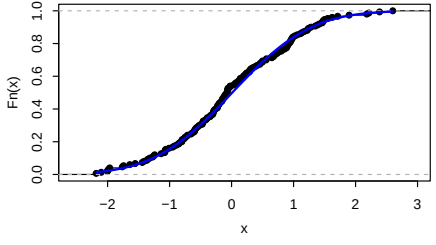


Illustration 1

échantillon-Normale







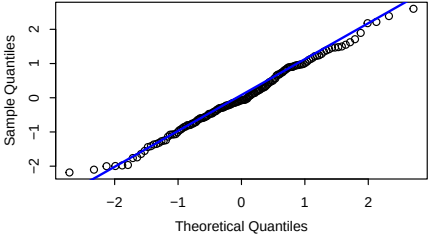
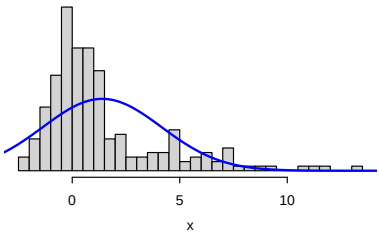
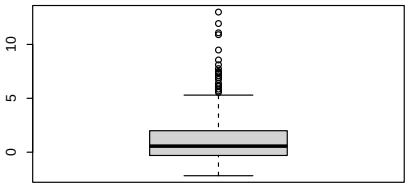
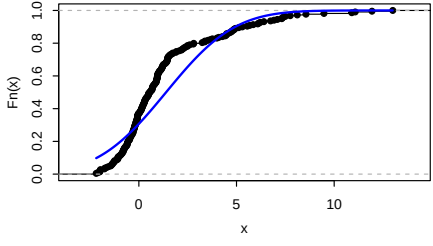


Illustration 2

échantillon-Asymétrique droite







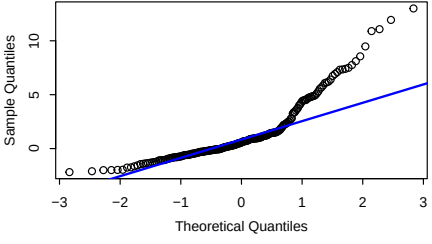
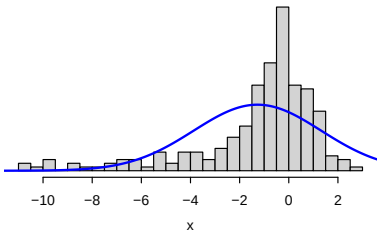
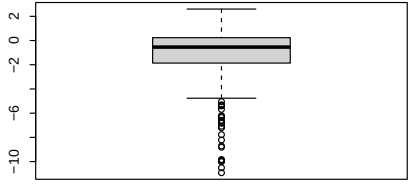
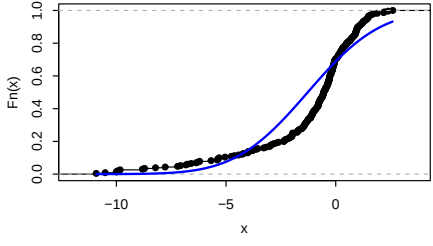


Illustration 3

échantillon-Asymétrique gauche







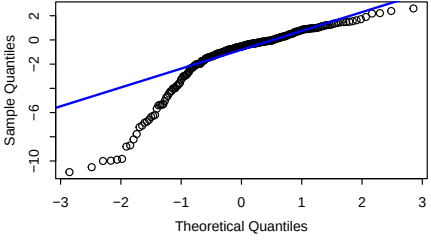
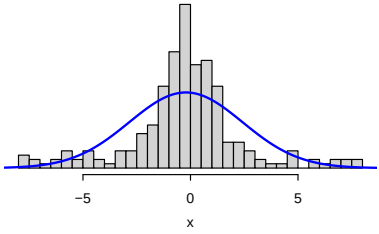
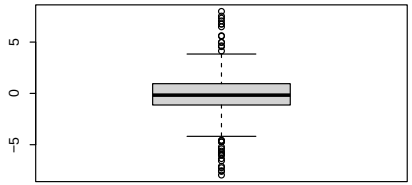
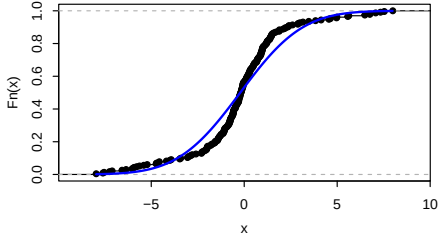


Illustration 4

échantillon-Peaked in middle







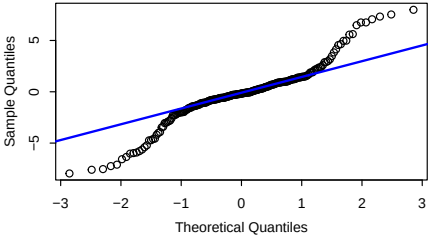
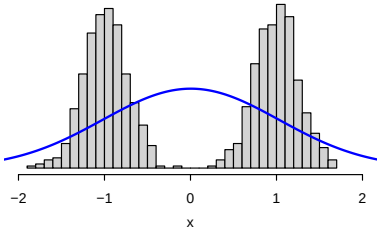
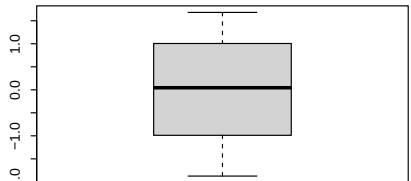
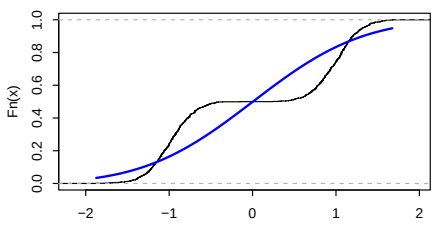


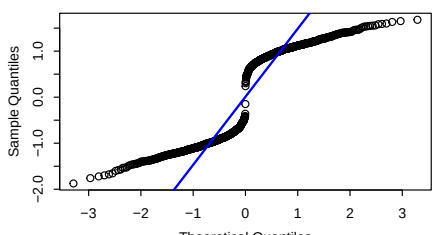
Illustration 5

échantillon-Bimodal









Commandes R les plus utiles vues dans ce chapitre

- dbinom, pbinom et rbinom
- dpois, ppois et rpois
- dunif, punif, qunif, et runif
- dnorm, pnorm, qnorm et rnorm
- plot, points, lines, et curve
- boxplot, hist, density, et ecdf
- qqnorm, qqline, et quantile
- replicate, sample
- table, proportions, mean