TP2 stats

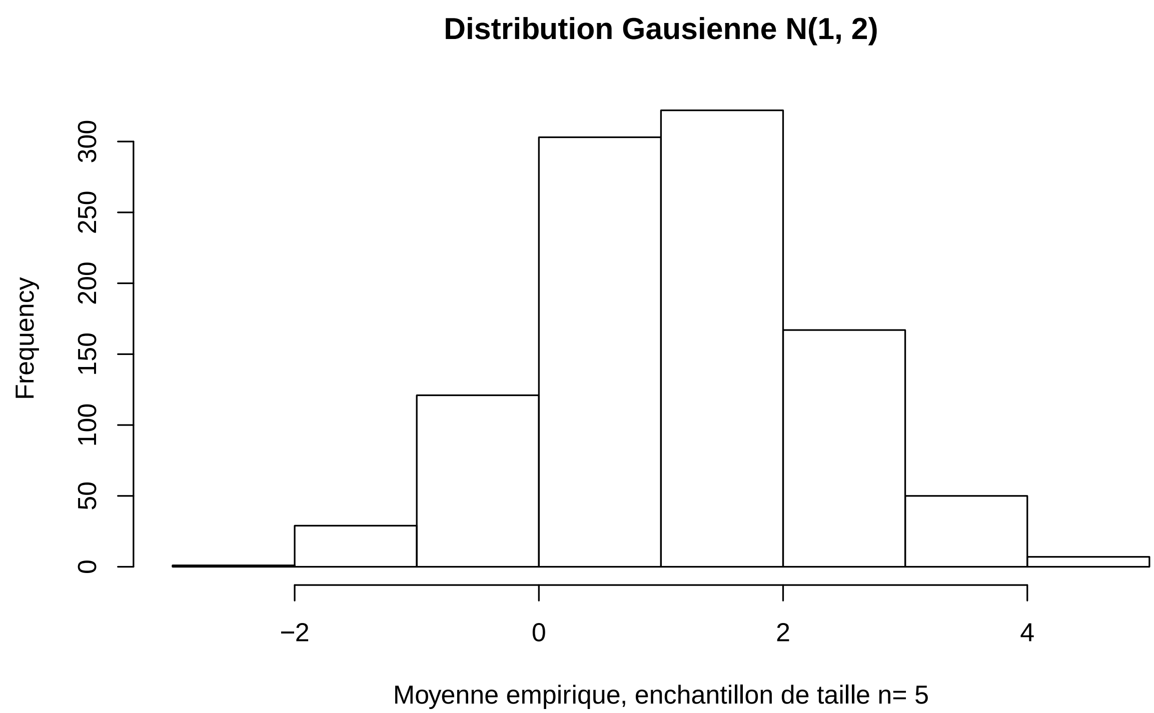
Romain PEREIRA

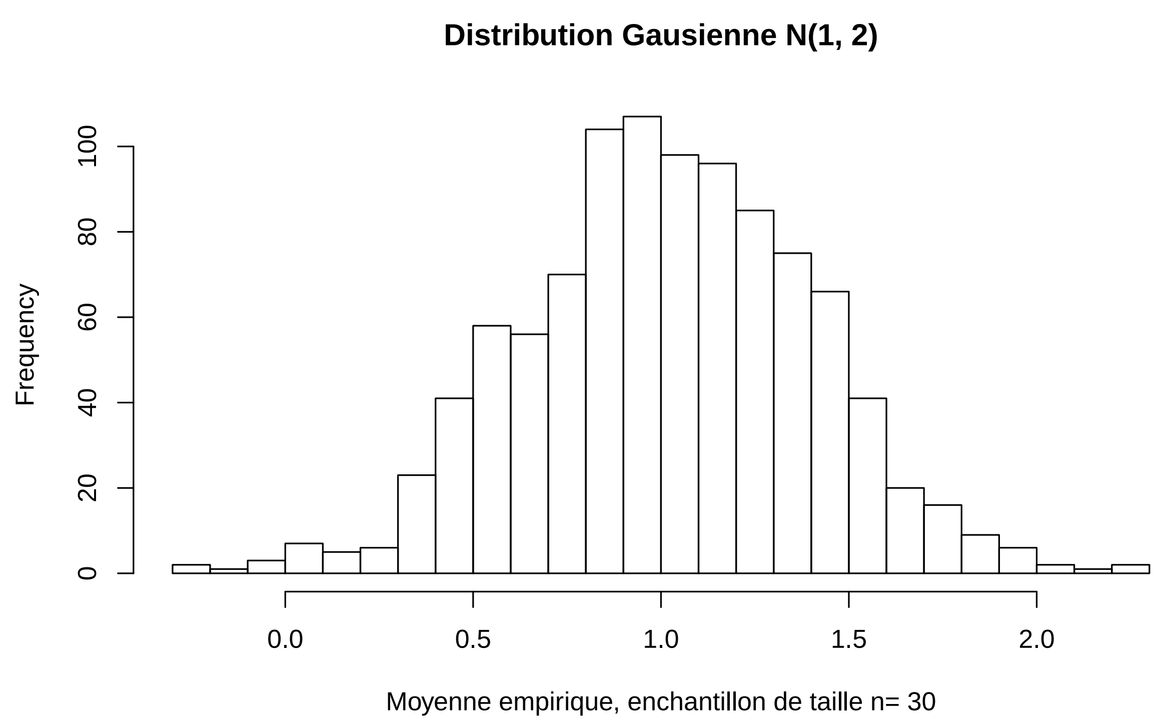
5 Mars 2018

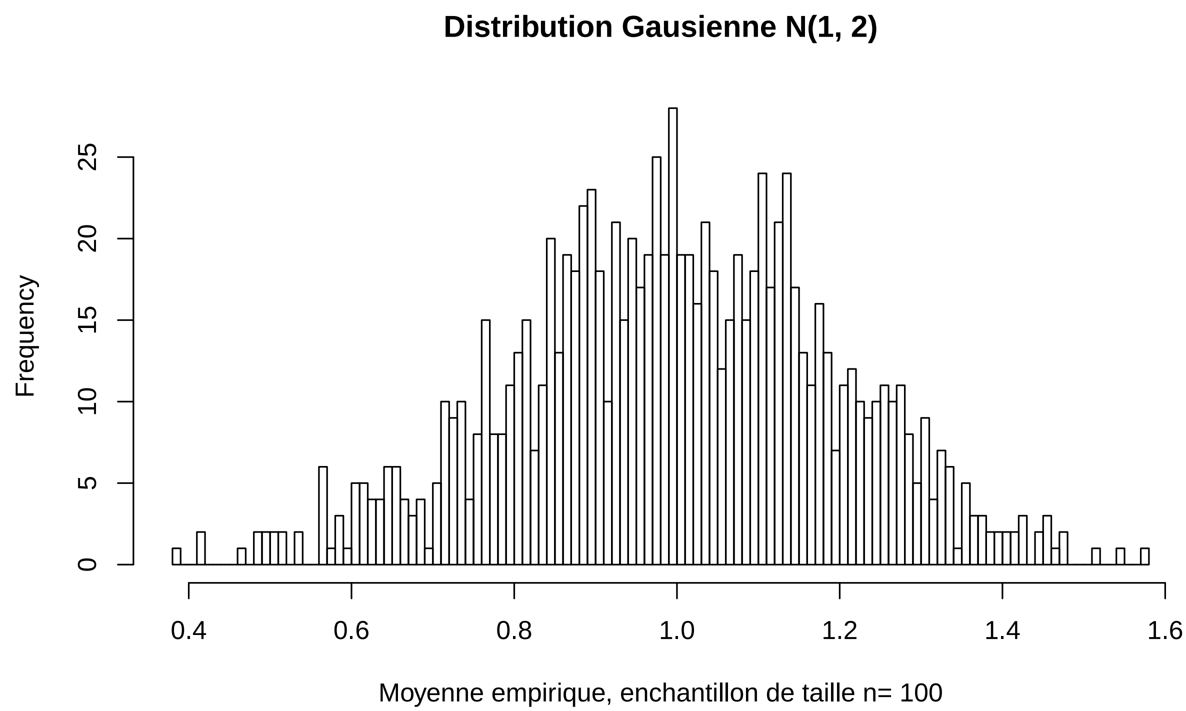
1. Echantillon, Théorème Central Limite, Estimation Monte Carlo

1.1 Simulation de 1000 échantillon i.i.d gaussien.

```r
N <- 1000
n <- c(5, 30, 100)

empirical_mean <- function(vec) {
  s <- 0
  for (x in vec) {
    s <- s + x
  }
  return (s / (length(vec) - 1))
}

empirical_var <- function(vec) {
  m <- empirical_mean(vec)
  s <- 0
  for (x in vec) {
    s <- s + (x - m) * (x - m)
  }
  return (s / (length(vec) - 1))
}

# fonction 'mean_hist'
#
# 'law' : fonction qui genere un vecteur de taille 'm', e.g: law(42)
# 'title': titre de l'histogramme
#
# La fonction trace 3 histogrammes de la loi de la moyenne empirique
# sur 'N' echantillons de taille dans 'ns'
mean_hist <- function(law, title) {
  for (nj in n) {
    sample <- law(nj * N)
    means <- c()
    for (i in 1:N) {
      subsample <- sample[((i-1)*nj + 1): (i * nj)]
      Xni <- empirical_mean(subsample)
      # vni <- empirical_var(subsample)
      means <- c(means, Xni)
    }
    hist(means, xlab=paste("Moyenne empirique, enchantillon de taille n=", nj), main=title, breaks=nj)
  }
}

mean_hist(function(n) { return (rnorm(n, mean=1, sd=2)) }, "Distribution Gausienne N(1, 2)")
```


Distribution Gausienne N(1, 2)
Frequency
0
50
100
150
200
250
300
−2
0
2
4
Moyenne empirique, enchantillon de taille n= 5


Distribution Gausienne N(1, 2)
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
Moyenne empirique, enchantillon de taille n= 30



Je pose $S_n = \sum_{i=1}^n X_i$, tel que:: + les X_i i.i.d de même loi + $E[X_i] = \mu$ + $\mathbb{V}[X_i] = \sigma^2$.

D'après le théorème central limite, pour N assez grand, S_n suit approximativement une loi normal $N(n\mu, n\sigma^2)$.

En notant la moyenne empirique $\bar{X}_n = \frac{S_n}{n}$, on a:

$\mathbb{E}[\bar{X}_n] = \mathbb{E}[\frac{S_n}{n}] = \frac{1}{n}\mathbb{E}[S_n] = \mu$

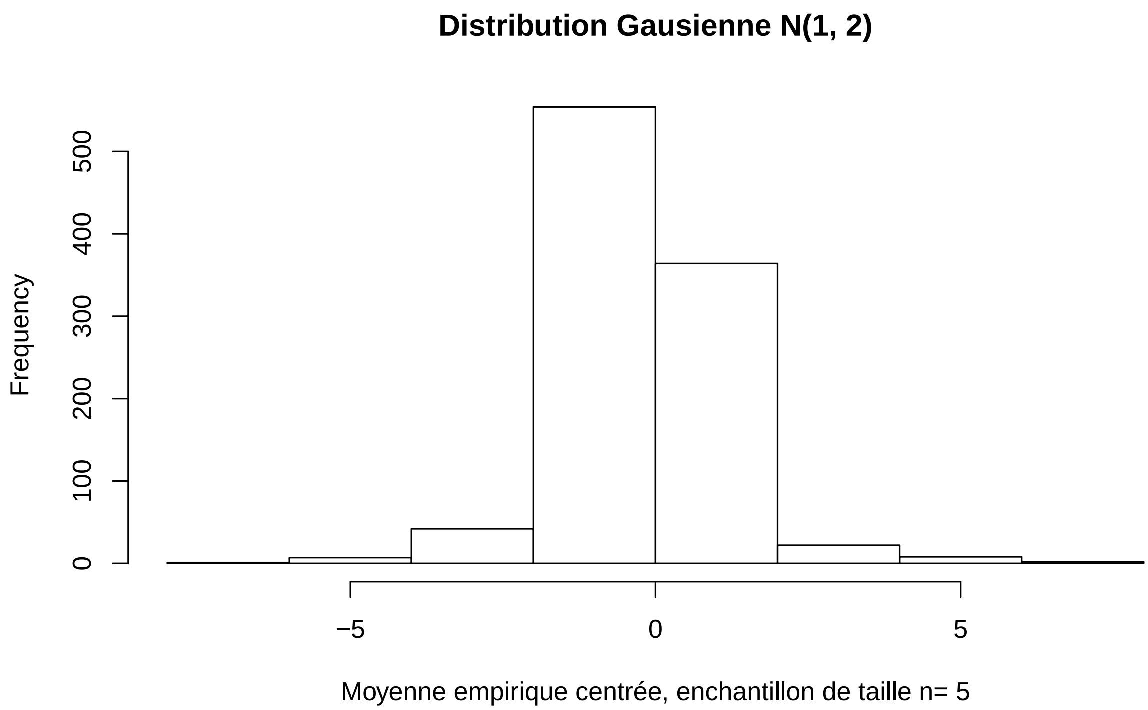
$\mathbb{E}[\bar{X}_n] = \mathbb{V}[\frac{S_n}{n}] = \frac{1}{n^2}\mathbb{V}[S_n] = \frac{\sigma^2}{n}$

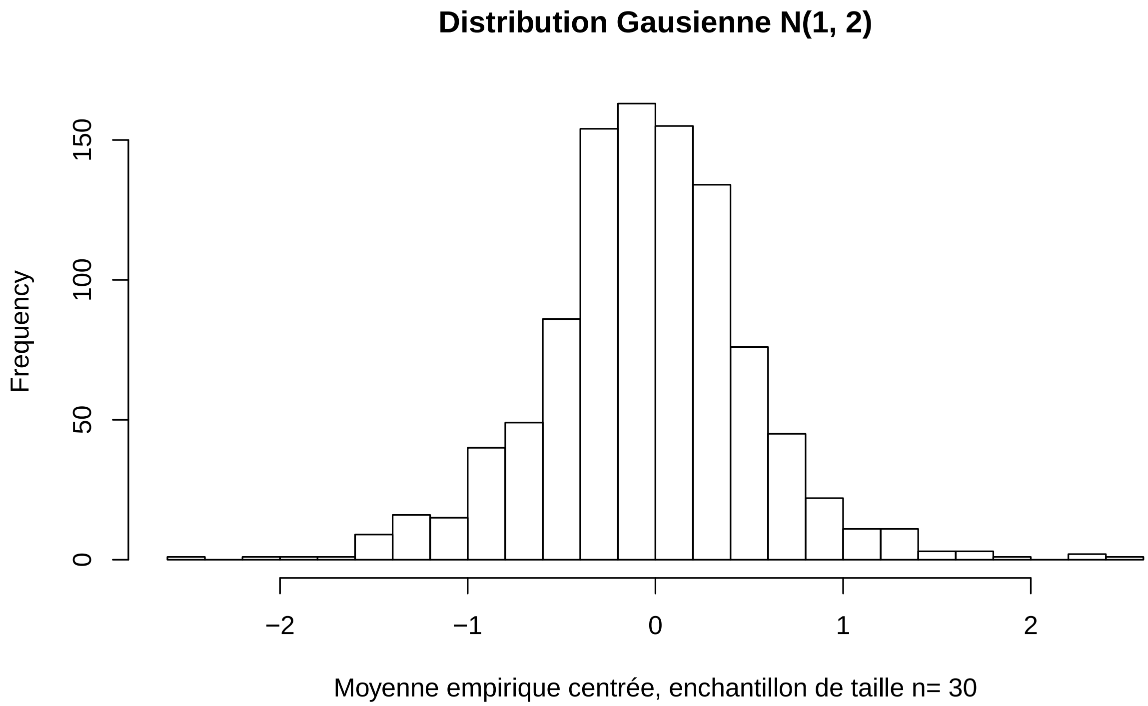
Dans notre exemple, les X_i suivent une loi $N(1, 2)$.

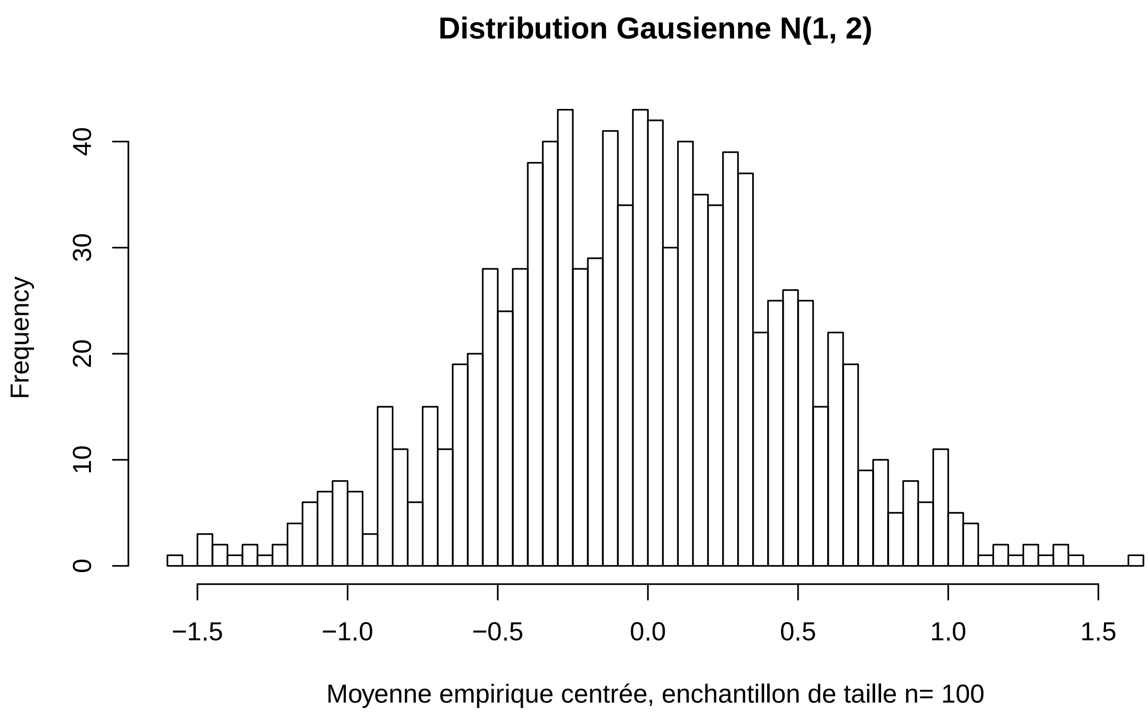
En notant $(a_n, b_n) = (\text{moyenne}, \text{ecart-type}) = (\mu, \sqrt{\frac{\sigma^2}{n}}) = (1, \frac{2}{\sqrt{n}})$, $U_n = \frac{\bar{X}_n - a_n}{b_n}$ suit une loi normal centrée réduite $N(0, 1)$.

```r
# fonction 'mean_norm_hist'
#
# 'law' : fonction qui genere un vecteur de taille 'm', e.g: law(42)
# 'title': titre de l'histogramme
#
# La fonction trace 3 histogrammes de la loi de la moyenne empirique renormalisé
# sur 'N' echantillons de taille dans 'ns'
mean_norm_hist <- function(law, title) {
  for (nj in n) {
    sample <- law(nj * N)
    Xn <- empirical_mean(sample)
    Un <- c()
    for (i in 1:N) {
      subsample <- sample[((i-1)*nj + 1): (i * nj)]
      ani       <- empirical_mean(subsample)
      bni       <- empirical_var(subsample) / sqrt(nj)
```

```
      Uni        <- (Xn - ani) / bni
      Un         <- c(Un, Uni)
    }
    hist(Un, xlab=paste("Moyenne empirique centrée, enchantillon de taille n=", nj), main=title, breaks=
  }
}

mean_norm_hist(function(n) { return (rnorm(n, mean=1, sd=2)) }, "Distribution Gausienne N(1, 2)")
```







Les histogrammes obtenus montrent en effet une loi normale centrée réduite.

Plus n est grand, plus la loi moyenne empirique renormalisé semble suivre une loi N(0, 1). (cf Théorème Central Limite)

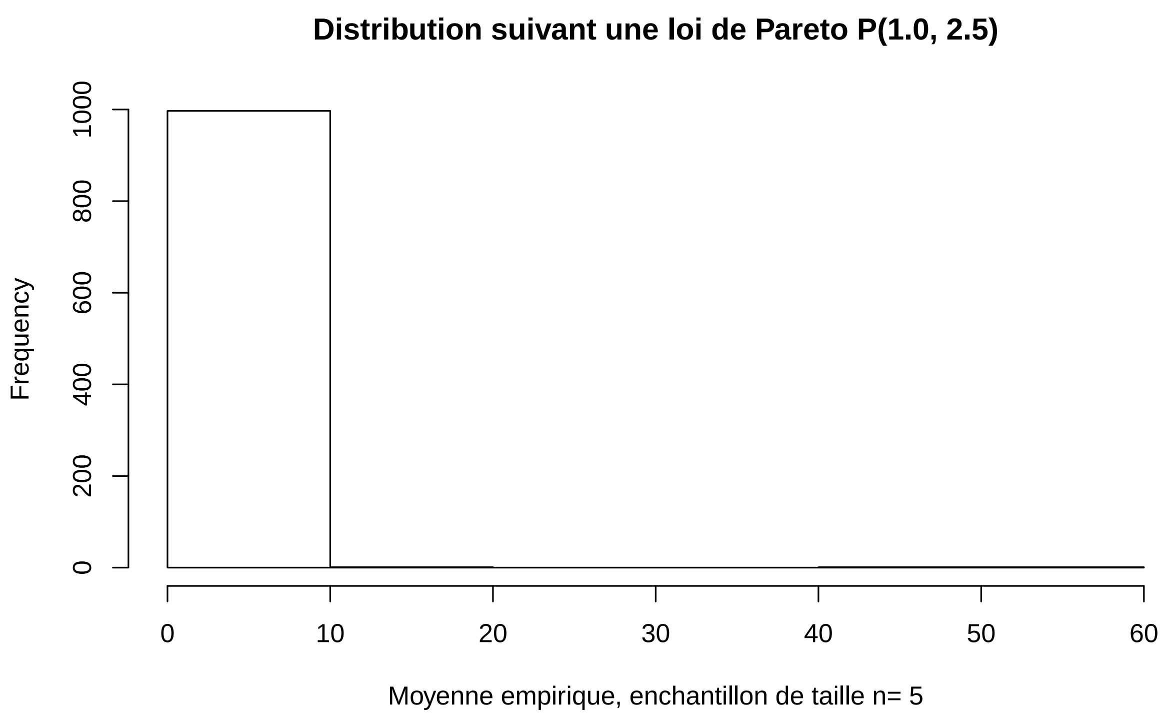
1.2 Loi de Pareto

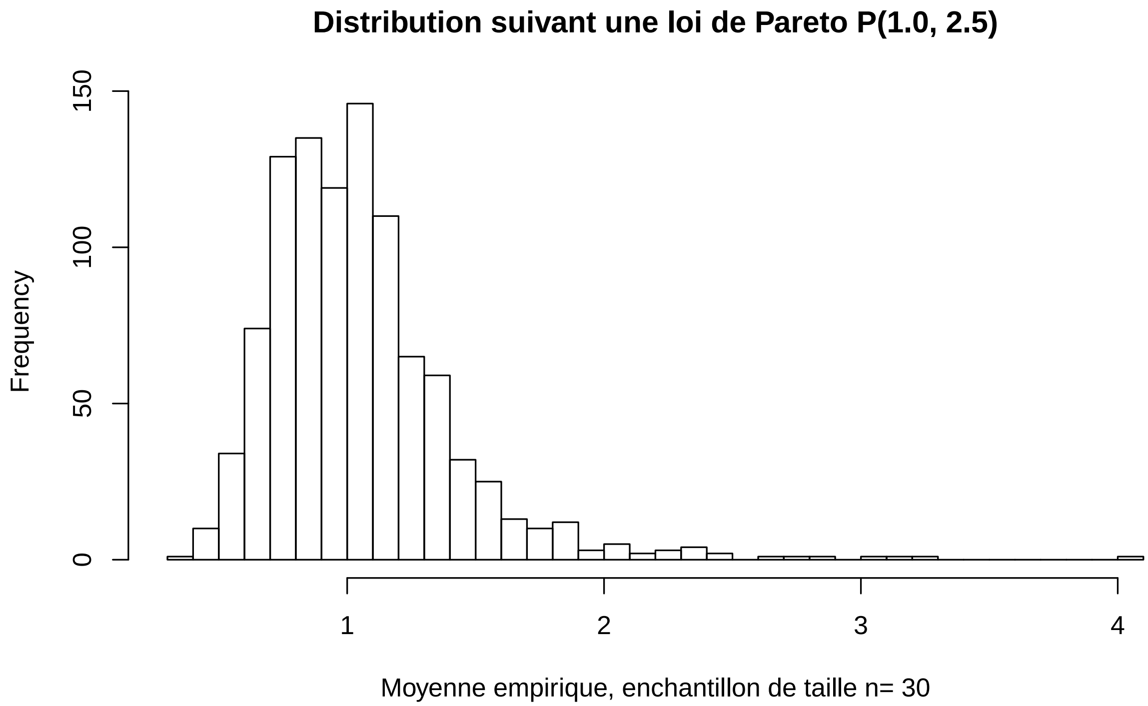
Soit X une variable aléatoire suivant une loi de Pareto $P(a, \alpha), \alpha > 2$. Alors, $\mathbb{E}[X] = \frac{\alpha a}{\alpha - 1}$ et $\mathbb{V}[X] = (\frac{\alpha a}{\alpha-1})^2 \frac{\alpha}{\alpha-2}$

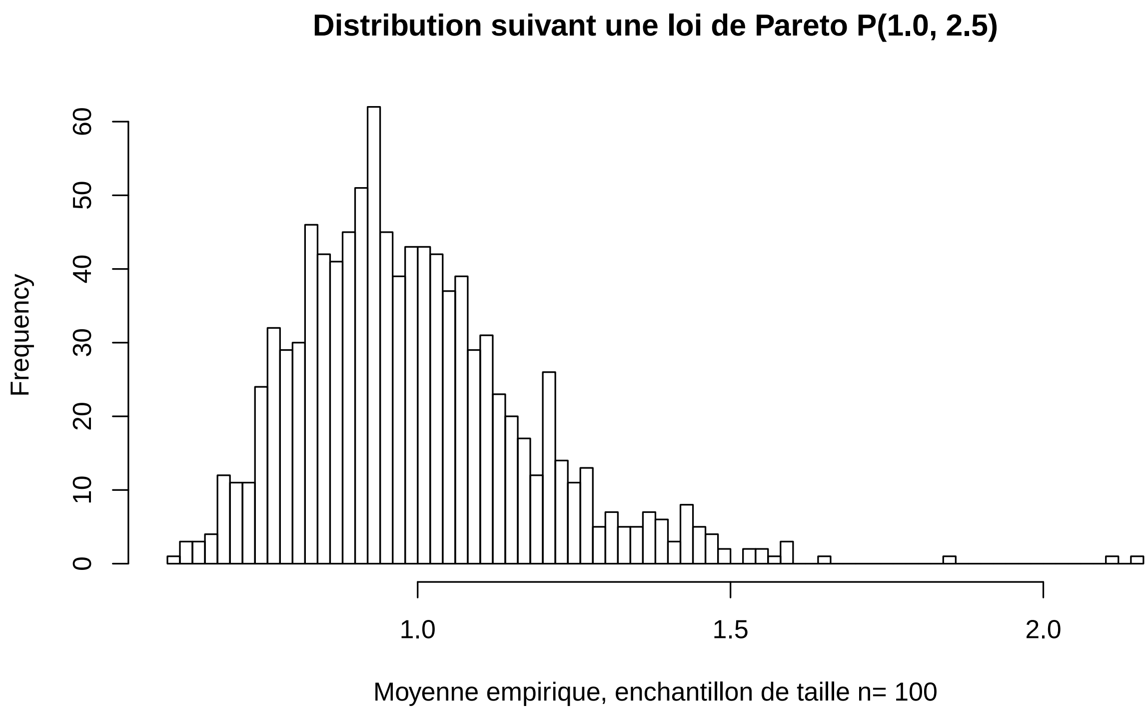
```
library("rmutil")
```

```
##
## Attaching package: 'rmutil'

## The following object is masked from 'package:stats':
##
##     nobs
```

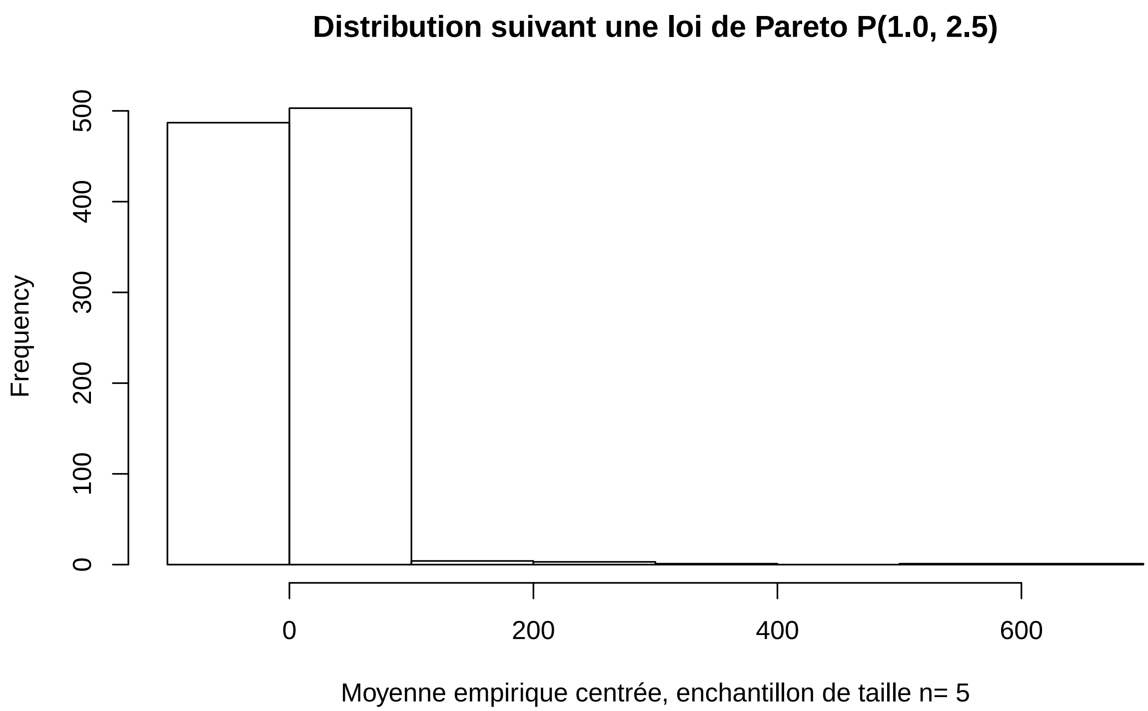
```
a     <- 1.0
alpha <- 2.5
mean_hist(function(n) { return (rpareto(n, m=a, s=alpha)) }, "Distribution suivant une loi de Pareto P(
```

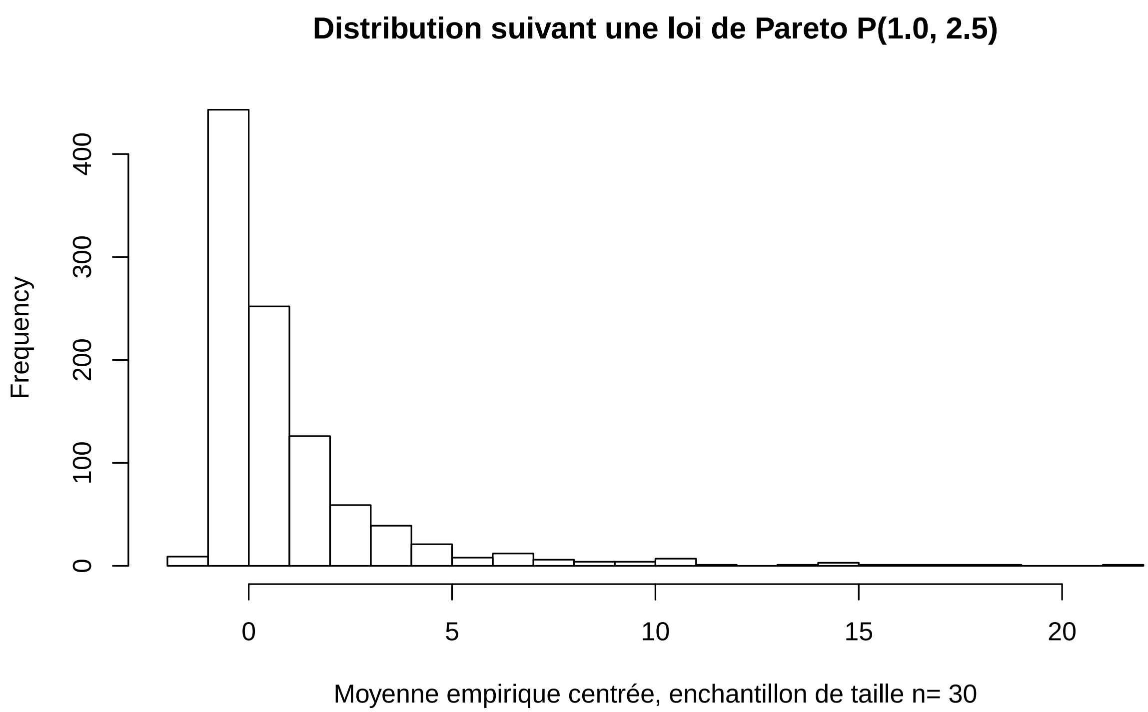


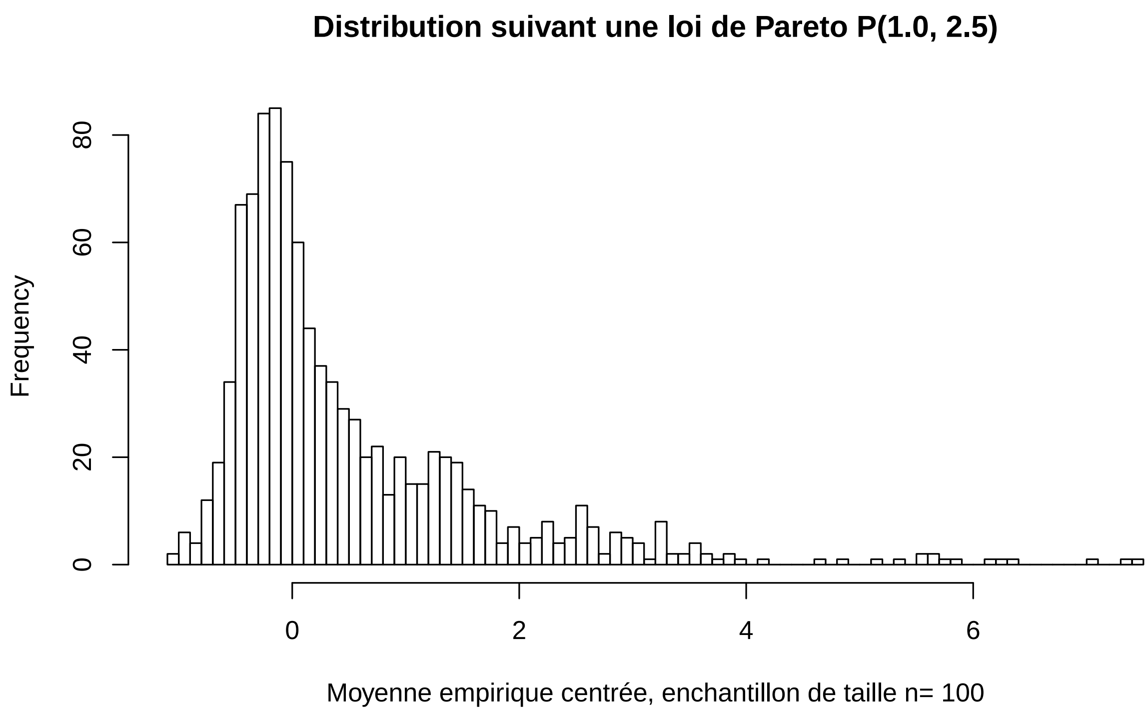





```
mean_norm_hist(function(n) { return (rpareto(n, m=a, s=alpha)) }, "Distribution suivant une loi de Pare
```


Distribution suivant une loi de Pareto P(1.0, 2.5)
Frequency
0
100
200
300
400
500
0
200
400
600
Moyenne empirique centrée, enchantillon de taille n= 5


Distribution suivant une loi de Pareto P(1.0, 2.5)
Frequency
0
100
200
300
400
0
5
10
15
20
Moyenne empirique centrée, enchantillon de taille n= 30

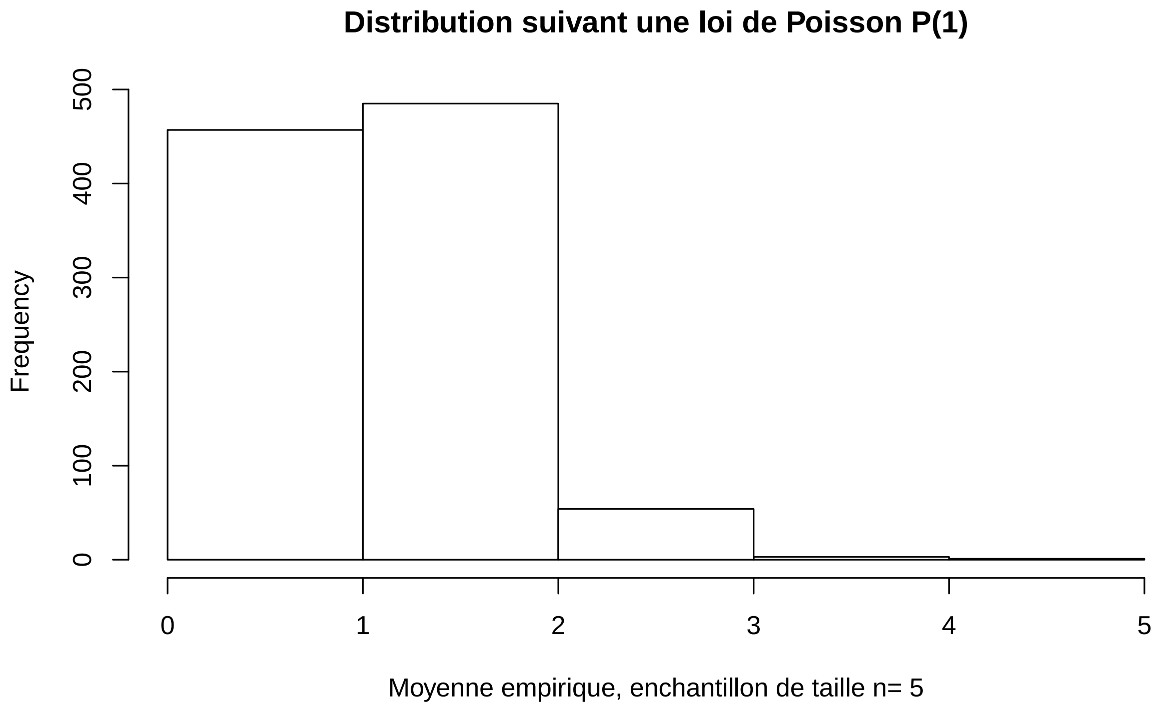


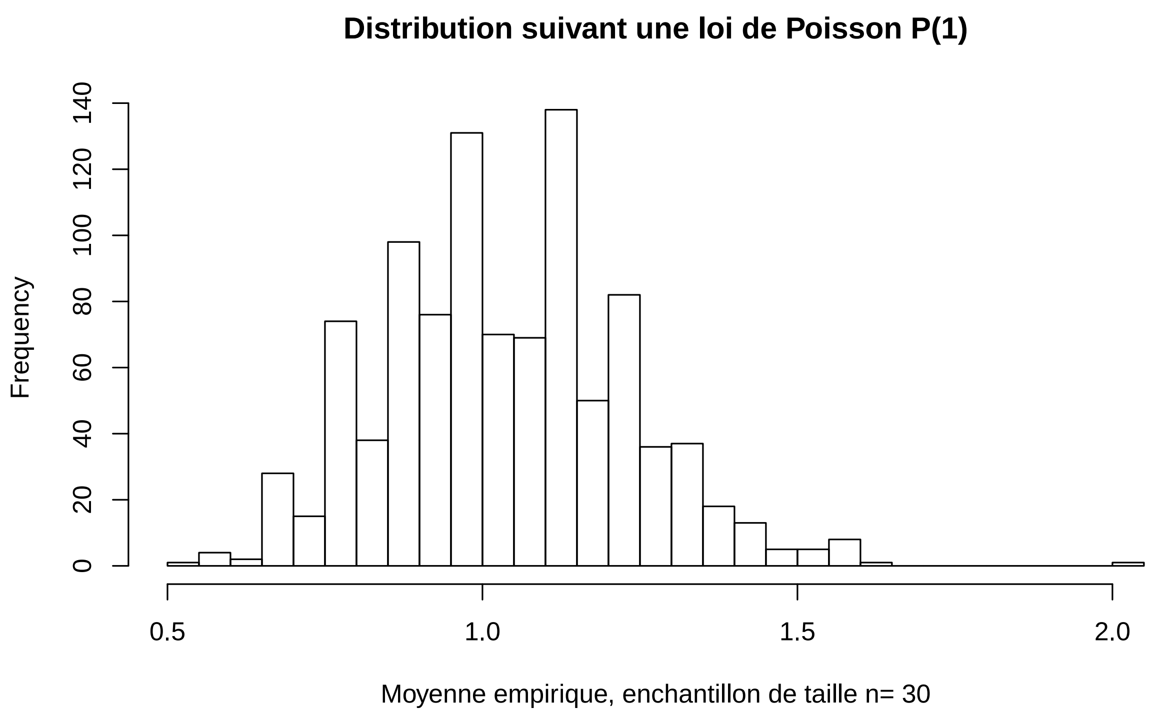
Plus n est grand, plus la loi moyenne empirique renormalisé semble suivre une loi N(0, 1) (bien qu'elle possède de nombreuses valeurs 'à droite' faisant penser à une allure exponentielle).

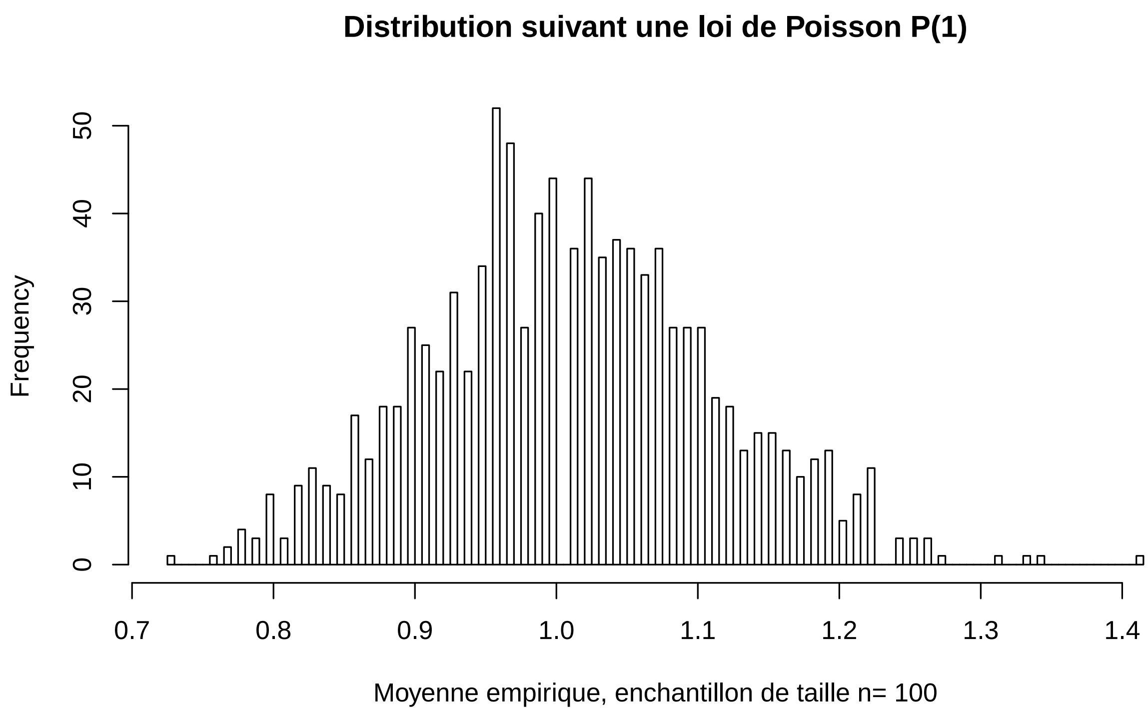
1.3 Loi de Poisson

Soit X une variable aléatoire suivant une loi de Poisson $P(\lambda)$. Alors, $\mathbb{E}[X] = \lambda$ et $\mathbb{V}[X] = \lambda$

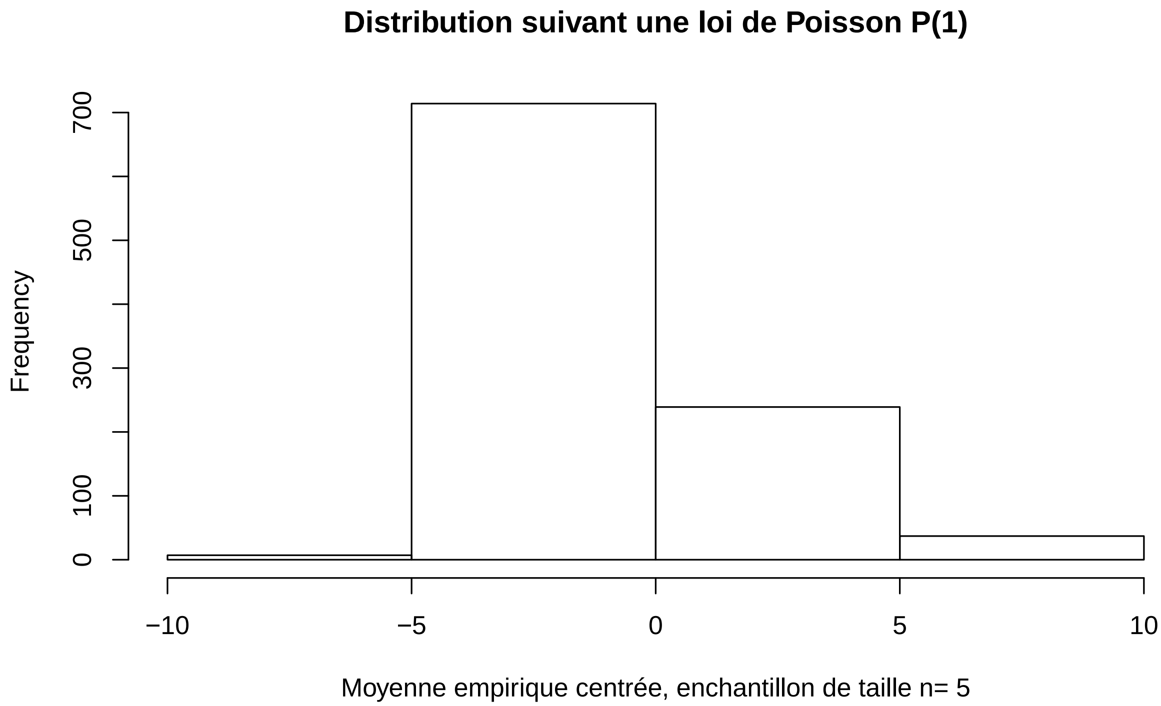
```
mean_hist(function(n) { return (rpois(n, lambda=1)) }, "Distribution suivant une loi de Poisson P(1)")
```

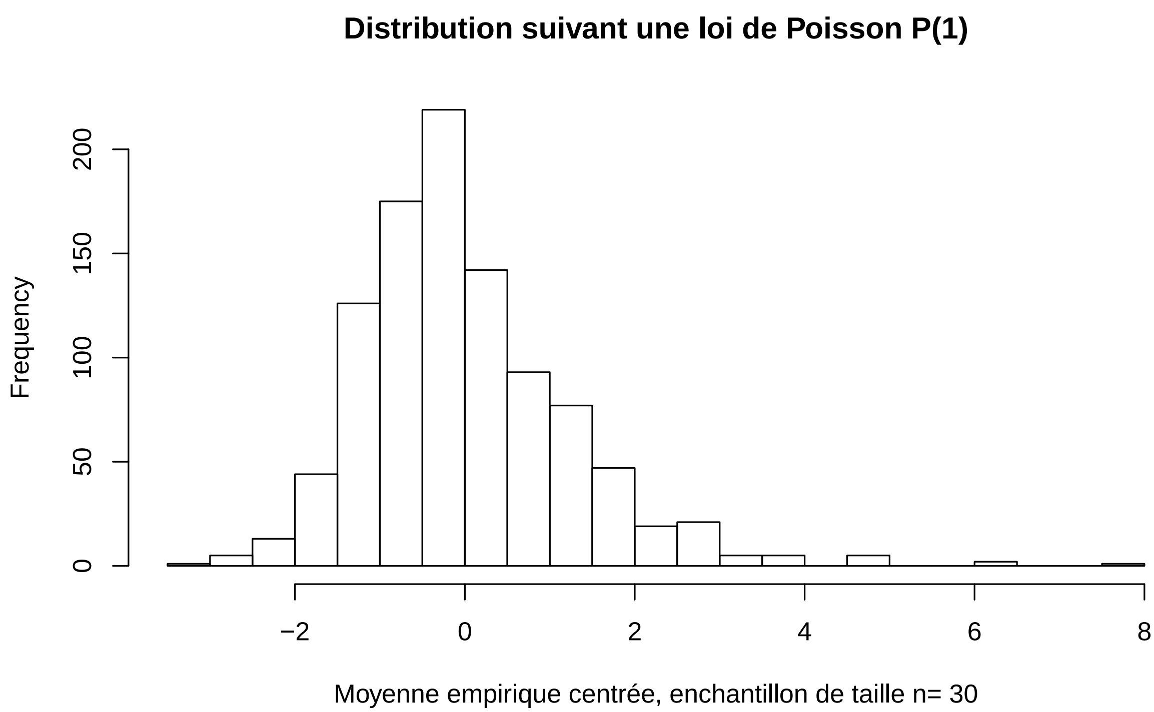

Distribution suivant une loi de Poisson P(1)
Frequency
0
100
200
300
400
500
0
1
2
3
4
5
Moyenne empirique, enchantillon de taille n= 5

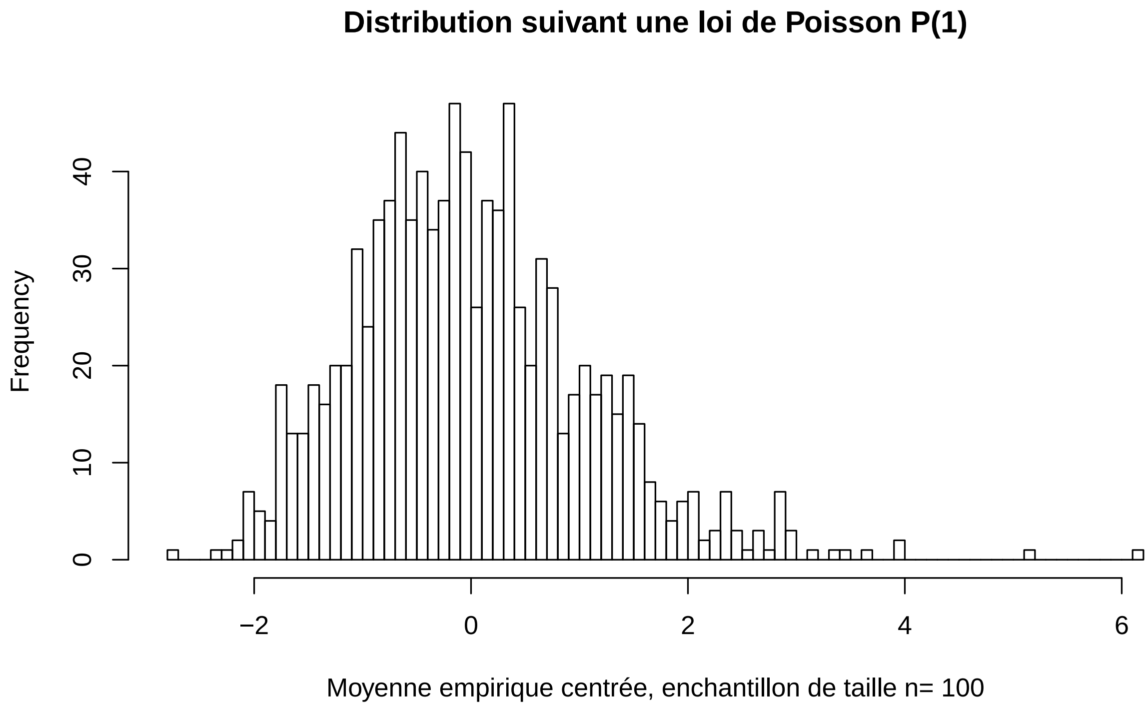

Distribution suivant une loi de Poisson P(1)
Frequency
0
20
40
60
80
100
120
140
0.5
1.0
1.5
2.0
Moyenne empirique, enchantillon de taille n= 30




```
mean_norm_hist(function(n) { return (rpois(n, lambda=1)) }, "Distribution suivant une loi de Poisson P(
```


Distribution suivant une loi de Poisson P(1)
Frequency
0
100
300
500
700
−10
−5
0
5
10
Moyenne empirique centrée, enchantillon de taille n= 5


Distribution suivant une loi de Poisson P(1)
Frequency
0
50
100
150
200
−2
0
2
4
6
8
Moyenne empirique centrée, enchantillon de taille n= 30



De même, plus n est grand, plus la loi moyenne empirique renormalisé semble suivre une loi N(0, 1).

1.4 Méthode d'expérimentation

On note $X = (X_1, ..., X_n)$ pour $n \in \mathbb{R}$, un échantillon de taille n (simulable 'facilement')

On suppose que tous les X_i son i.i.d, et suivent la même loi.

Soit $T : \Omega^n \to \mathbb{R}$ une statistique sur un echantillon de taille n.

On peut trouver une approximation de l'espérance $\mathbb{E}[T(X)]$ en utilisant le protocole suivant:

1. Fixer $N \in \mathbb{N}, N \gg 1$.
2. Générer N échantillons de taille n, notés $X^i = (X_1^i, ..., X_n^i)$ avec $1 \leq i \leq N$
3. Je pose

- $\bar{T}_N = \frac{1}{N} \sum_{i=1}^{N} T(X^i)$
- Lorsque N devient grand, d'après le théorème central limite (même raisonnement qu'en **1.1**):
  - (1) : $\mathbb{E}[\bar{T}_N] \xrightarrow[N \gg 1]{} \mathbb{E}[T(X)]$
  - (2) : $\mathbb{V}[\bar{T}_N] \xrightarrow[N \gg 1]{} \frac{1}{N} \mathbb{V}[T(X)] \xrightarrow[N \to +\infty]{} 0$

4. Donc d'après (1) et (2), on a $\bar{T}_n \xrightarrow[N \to +\infty]{\mathbb{L}^2} \mathbb{E}[T(X)] = cste$.

Autrement dit, la moyenne empirique (entant que variable aléatoire), tends (en norme 2) vers la v.a constante $\mathbb{E}[T(X)]$.

La moyenne empirique est une bonne estimation de l'espérance quand elle est effectué sur un grand nombre d'échantillon.

2. Moyenne et dispersion.

2.1 Inégalité de Bienaymé Tchebychev

Soit X une variable aléatoire admettant un moment d'ordre 2 (et donc un moment d'ordre 1). L'inégalité de Bienaymé-Tchebychev affirme: $\forall \alpha \in \mathbb{R}_+^*, \mathbb{P}(|X - \mathbb{E}[X]| \geq \alpha) \leq \frac{1}{\alpha^2}\mathbb{V}[X]$

Pour une loi Gaussienne $N(\mu, \sigma^2)$, on a : $\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$

Pour une loi de Poisson $P(\lambda)$, on a : $\mathbb{P}(|X - \lambda| \geq \alpha) \leq \frac{\lambda}{\alpha^2}$

2.2 Estimation par Monte Carlo.

(a) $\mathbb{P}(|X - \mu| \geq \delta) = \mathbb{E}[1_{\{|X-\mu|\geq\delta\}}] = \mathbb{E}[Z]$, en posant $Z = 1_{\{|X-\mu|\geq\delta\}}$

(b) On estime $\mathbb{E}[Z]$ par la moyenne empirique $\bar{Z}_N = \frac{1}{N}\sum_{i=1}^{N} T(Z^i)$:

```r
# effectue une estimation de Monte Carlo sur une loi Gaussienne, de Pareto, et de Poisson.
#
# N : nombre d'echantillon pour la moyenne
# delta : verifiant (a)
# (mu, sigma) : paramètre de la loi Gausienne
# (a, alpha) : paramètre de la loi de Pareto
# (lambda) : paramètre de la loi de Poisson
#
# renvoie une liste contenant :
#           - les distributions générées
#           - la moyenne empirique de ces distributions
#           - les distributions transformées Z (voir (a))
#           - les espérances empirique de Z, approximation de (a)
estimate_monte_carlo <- function(N, delta, mu, sigma, a, alpha, lambda) {
  # On genere des distributions
  XN       <- list("Gauss" = rnorm(N, mu, sigma), "Pareto" = rpareto(N, a, alpha),
                   "Poisson" = rpois(N, lambda))
  XN_means <- list("Gauss" = mu,                  "Pareto" = alpha*a/(alpha - 1),
                   "Poisson" = lambda)
  ZN       <- list()
  ZN_means <- list()
  # Pour chaque distributions
  for (distrib in names(XN)) {
    # on recupere la distribution
    XNi      <- XN[[distrib]]
    XNi_mean <- XN_means[[distrib]]
    # on génère la variable aléatoire Z correspondante
    ZN[[distrib]] <- unlist(lapply(XNi, function(xi) {
                              if (abs(xi-XNi_mean) >= delta) {
                                return (1)
                              }
                              return (0)
                            }))
    ZN_means[[distrib]] <- empirical_mean(ZN[[distrib]])
  }
  return (list("XN" = XN, "XN_means" = XN_means, "ZN" = ZN, "ZN_means" = ZN_means))
}
```

On obtient selon les différentes lois: $\mathbb{P}(|X - \mathbb{E}[X]| \geq \delta_{=1}) = ...$

```
N <- 1e5
estimation <- estimate_monte_carlo(N, delta=1, mu=0, sigma=1, a=1.0, alpha=2.5, lambda=1)
estimation[["ZN_means"]]
```

```
## $Gauss
## [1] 0.3178632
##
## $Pareto
## [1] 0.6788168
##
## $Poisson
## [1] 0.6332163
```

La moyenne empirique est une variable aléatoire, et on a montré que $\mathbb{E}[\bar{Z}_N] = \mathbb{E}[Z]$ et une variance $\mathbb{V}[\bar{Z}_N] = \frac{1}{N}\mathbb{V}[Z]$.

Donc d'après le théorème de Bienaymé Tchebichev, la précision de notre estimation $\mathbb{P}(|X - \mu| \geq \delta) = \mathbb{E}[Z] \simeq \bar{Z}_N$, est donné par:

$\forall \epsilon \geq 0, \mathbb{P}[|\bar{Z}_N - \mathbb{E}[Z]| \geq \epsilon] \leq \frac{1}{\epsilon N}\mathbb{V}[Z]$.

(c) Application numérique:

```
markov_sup <- function(Z, N, eps) {
  return (var(Z) / (eps * N));
}
eps <- 1e-4
```

On a fixé plus tôt $N = 10^5$. On fixe $\epsilon = 10^{-4}$.

En fonction des loi de X précèdentes, notre estimation de $\bar{Z}_N \simeq \mathbb{E}[Z]$ vérifie:

$\mathbb{P}[|\bar{Z}_N - \mathbb{E}[Z]| \geq \epsilon] = \mathbb{P}(\bar{Z}_N \notin [\mathbb{E}[Z] - \epsilon; \mathbb{E}[Z] + \epsilon]) \leq ...$

```
XN <- estimation[["XN"]]
ZN <- estimation[["ZN"]]
for (distrib in names(XN)) {
  print(paste(distrib, ":", markov_sup(ZN[[distrib]], N, eps)))
}
```

```
## [1] "Gauss : 0.0216827188671887"
## [1] "Pareto : 0.0218029164191642"
## [1] "Poisson : 0.0232257418474185"
```

Remarques:

- Plus ϵ est 'petit', plus notre précision est incertaine. (la probabilité que notre estimation soit dans l'invervalle $[\mathbb{E}[Z] - \epsilon; \mathbb{E}[Z] + \epsilon]$ s'éloigne de 1)
- Plus N est 'grand', plus notre précision est probable. (la probabilité que notre estimation soit dans l'invervalle $[\mathbb{E}[Z] - \epsilon; \mathbb{E}[Z] + \epsilon]$ tends vers 1)
- Plus $\mathbb{V}[Z]$ est 'grande', plus notre précision est incertaine. (la probabilité que notre estimation soit dans l'invervalle $[\mathbb{E}[Z] - \epsilon; \mathbb{E}[Z] + \epsilon]$ s'éloigne de 1)

Voici les bornes obtenus pour différentes valeurs de δ et σ:

```
print("Majoration par l'inégalité de Markov, de la probabilité que la moyenne obtenu s'écarte à +- epsi
```

```
## [1] "Majoration par l'inégalité de Markov, de la probabilité que la moyenne obtenu s'écarte à +- epsi
```

```r
for (delta in c(1, 5, 10)) {
  for (sigma in c(1, 10, 100)) {
    estimation <- estimate_monte_carlo(N, delta, mu=0, sigma, a=1.0, alpha=2.5, lambda=1)
    XN <- estimation[["XN"]]
    ZN <- estimation[["ZN"]]
    print("-----------------------------")
    print(paste("Pour delta=", delta, " et sigma=", sigma, sep=""))
    for (distrib in names(XN)) {
      print(paste(distrib, ":", markov_sup(ZN[[distrib]], N, eps)))
    }
  }
}
```

```
## [1] "-----------------------------"
## [1] "Pour delta=1 et sigma=1"
## [1] "Gauss : 0.0216608258482585"
## [1] "Pareto : 0.0217335225752258"
## [1] "Poisson : 0.023344379203792"
## [1] "-----------------------------"
## [1] "Pour delta=1 et sigma=10"
## [1] "Gauss : 0.00727849869498695"
## [1] "Pareto : 0.0219008587585876"
## [1] "Poisson : 0.0232875371853719"
## [1] "-----------------------------"
## [1] "Pour delta=1 et sigma=100"
## [1] "Gauss : 0.000806398373983739"
## [1] "Pareto : 0.0218467528275283"
## [1] "Poisson : 0.0232843948339483"
## [1] "-----------------------------"
## [1] "Pour delta=5 et sigma=1"
## [1] "Gauss : 0"
## [1] "Pareto : 0.00141636447364474"
## [1] "Poisson : 6.29609396093961e-05"
## [1] "-----------------------------"
## [1] "Pour delta=5 et sigma=10"
## [1] "Gauss : 0.0235777024870249"
## [1] "Pareto : 0.00144258146581466"
## [1] "Poisson : 6.3959679596796e-05"
## [1] "-----------------------------"
## [1] "Pour delta=5 et sigma=100"
## [1] "Gauss : 0.00391816693166932"
## [1] "Pareto : 0.00149788056880569"
## [1] "Poisson : 5.59691996919969e-05"
## [1] "-----------------------------"
## [1] "Pour delta=10 et sigma=1"
## [1] "Gauss : 0"
## [1] "Pareto : 0.000473738977389774"
## [1] "Poisson : 0"
## [1] "-----------------------------"
## [1] "Pour delta=10 et sigma=10"
## [1] "Gauss : 0.0217027242172422"
## [1] "Pareto : 0.00042220646206462"
## [1] "Poisson : 0"
## [1] "-----------------------------"
```

```
## [1] "Pour delta=10 et sigma=100"
## [1] "Gauss : 0.00735923358233583"
## [1] "Pareto : 0.00042319798197982"
## [1] "Poisson : 0"
```

(d) Inégalité de Chernoff.

Soit X une variable aléatoire admettant une fonction génératrice.

L'inégalité de Chernoff donne:

$\forall t \in [0,1], \forall \epsilon \in \mathbb{R}_+^*$,

$\mathbb{P}(X - \mathbb{E}[X] \geq \epsilon) \leq e^{-\epsilon t}\mathbb{E}[e^{(X-\mathbb{E}[X])t}]$

$\mathbb{P}(X - \mathbb{E}[X] \leq -\epsilon) \leq e^{-\epsilon t}\mathbb{E}[e^{(X-\mathbb{E}[X])t}]$

Donc,

$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2e^{-\epsilon t}\mathbb{E}[e^{(X-\mathbb{E}[X])t}]$

2.3.

2.4.

(a)

```r
theta <- 0
for (n in c(20, 100, 1000, 10000)) {
  cauchy <- rcauchy(n, location=theta, scale=1)
  m      <- empirical_mean(cauchy)
  print(paste("n=", n, " ; la moyenne empirique calculé est: ", m, sep=""))
}
```

```
## [1] "n=20 ; la moyenne empirique calculé est: -4.23762419760951"
## [1] "n=100 ; la moyenne empirique calculé est: -0.972571748183454"
## [1] "n=1000 ; la moyenne empirique calculé est: 9.15669176361602"
## [1] "n=10000 ; la moyenne empirique calculé est: -0.22214337262099"
```

La moyenne empirique donne des valeurs très différentes selon 'n', et ne semble pas converger.

(b) Une variable aléatoire X suivant une loi de Cauchy $C(\theta)$ n'admet pas d'espérance:

$f_X(x,\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$, et quand $x \to +\infty$, $xf_X(x,\theta) \sim \frac{1}{x}$, donc:

$\mathbb{E}[X] = \int_{-\infty}^{+\infty} |xf_X(x,\theta)|dx$ diverge.

Donc le théorème central limite ne s'applique pas: il n'y a pas d'espérance, donc la moyenne empirique ne converge pas.

Ceci s'explique par le fait que la probabilité d'obtenir une valeur éloigné de θ (la médiane) est trop elévé pour que la moyenne converge.

(c) La médiane d'une loi de Cauchy $C(\theta)$ est θ.

Si l'on sait qu'une phénomène suit une loi de Cauchy, il est possible de déterminer son paramètre θ en suivant ce protocole:

1. Fixer $n \in \mathbb{N}, n \gg 1$.
2. Générer un échantillon de taille n.
3. Trier les valeurs de cette échantillon par ordre croissant. (ou décroissant)
4. La valeur au centre de l'échantillon trié (en $\frac{n}{2}$) est un estimateur de θ.

Application:

```r
theta <- 0
for (theta in c(-1, 0, 1)) {
  print("-------------------------------------------------------")
  print(paste("theta=", theta, sep=""))
  for (n in c(20, 100, 1000, 10000)) {
    cauchy <- rcauchy(n, location=theta, scale=1)
    sorted <- sort(cauchy)
    print(paste("la médiane de l'échantillon n=", n, " vaut:", sorted[n / 2 + 1], sep=""))
  }
}
```

```
## [1] "-------------------------------------------------------"
## [1] "theta=-1"
## [1] "la médiane de l'échantillon n=20 vaut:-1.05030140332101"
## [1] "la médiane de l'échantillon n=100 vaut:-0.958551332956296"
## [1] "la médiane de l'échantillon n=1000 vaut:-0.9932978528048"
## [1] "la médiane de l'échantillon n=10000 vaut:-1.0008813125498"
## [1] "-------------------------------------------------------"
## [1] "theta=0"
## [1] "la médiane de l'échantillon n=20 vaut:0.138786591402866"
## [1] "la médiane de l'échantillon n=100 vaut:-0.276566051881372"
## [1] "la médiane de l'échantillon n=1000 vaut:-0.00221097533284202"
## [1] "la médiane de l'échantillon n=10000 vaut:0.0237195748443169"
## [1] "-------------------------------------------------------"
## [1] "theta=1"
## [1] "la médiane de l'échantillon n=20 vaut:0.779489512034759"
## [1] "la médiane de l'échantillon n=100 vaut:1.00727968868033"
## [1] "la médiane de l'échantillon n=1000 vaut:0.999919145980501"
## [1] "la médiane de l'échantillon n=10000 vaut:0.993892135005832"
```

Les valeurs obtenus par la simulation sont en accord avec celle attendu par notre protocole.