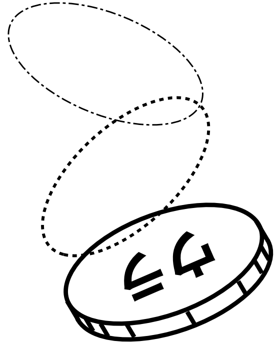
◆ コインとサイコロ，そして，大数の法則

コインを投げて，表の出る比率を考える．

... 経験から，まずは，想像して下さい ...



期待値「0.5」に対して，**実現値**はサマザマ,...

例えば，１０回中３回表がでた場合，

... 表の出る比率（実現値）は，3/10 = 0.3

つまり， | ある実現値は，１つの標本値であり，確率変数である．

具体的に実験してみよう．

ここでは，モンテカルロシミュレーションを行う．

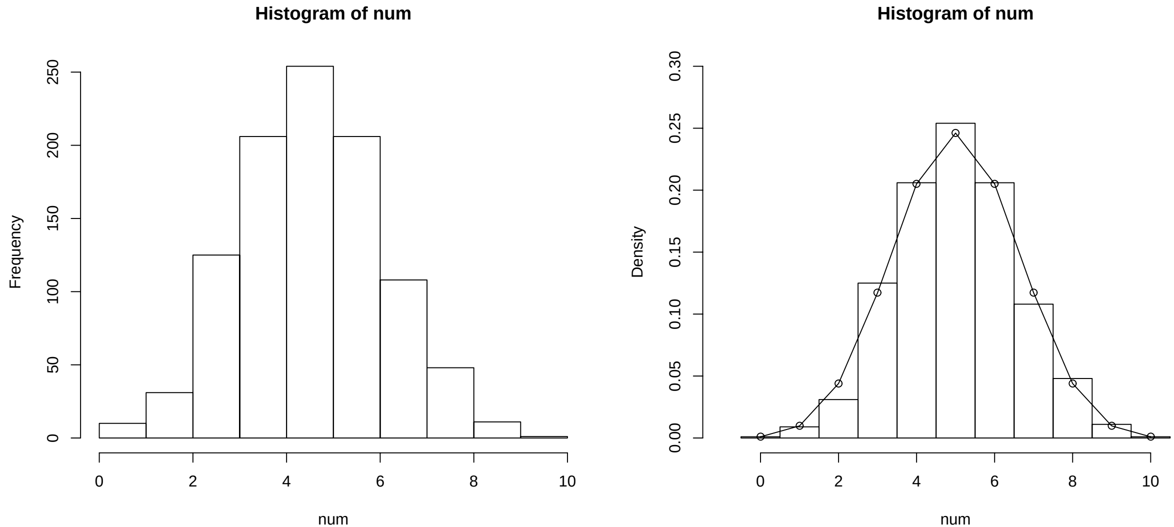
(コンピュータにコインを投げさせる ＝ 確率変動をコントロールできる．

実際に，人間がコインを投げる場合との違いを考えよ.）

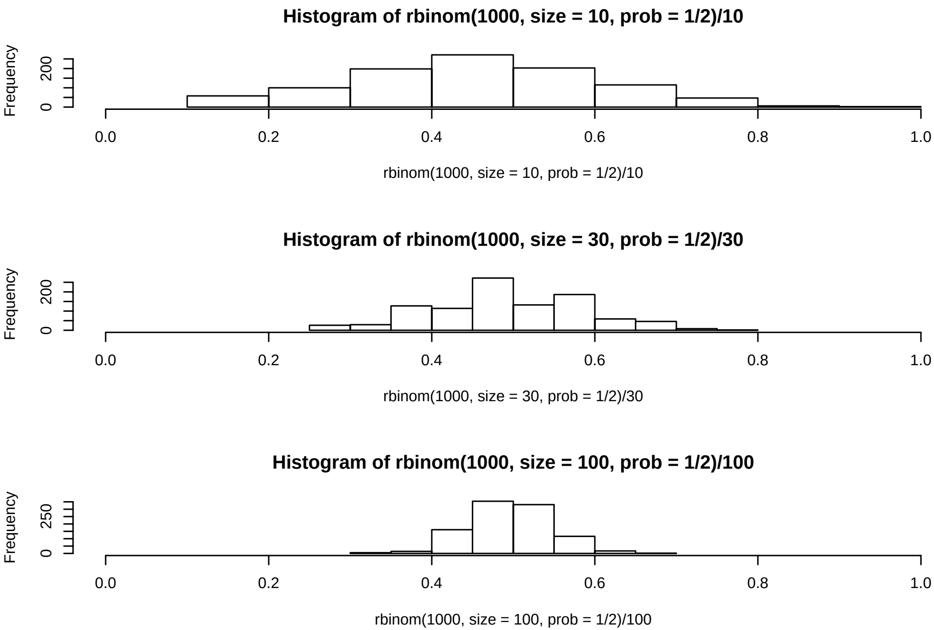
```
> # coin
> sample(c("omote","ura"), 1)
[1] "omote"
> sample(c("omote","ura"), 5, replace=TRUE)
[1] "omote" "ura"   "ura"   "omote" "ura"
>
> rbinom(1, size=1, prob=1/2) # 1 stands for “omote”; 0 for “ura”.
[1] 0
> rbinom(5, size=1, prob=1/2)
[1] 1 0 1 0 0
> sum(rbinom(5, size=1, prob=1/2))
[1] 2
> rbinom(1, size=5, prob=1/2)
[1] 3
>
> rbinom(20, size=5, prob=1/2)
 [1] 3 4 3 3 4 3 3 2 1 2 3 3 4 2 2 4 2 2 3 4
> dbinom(0:10, size=10, prob=1/2)
 [1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000 0.2050781250
 [6] 0.2460937500 0.2050781250 0.1171875000 0.0439453125 0.0097656250
[11] 0.0009765625
```

コインのオモテの数は
２項分布でモデル化できる

(3)




```
> table(num <- rbinom(1000, size=10, prob=1/2))
  0   1   2   3   4   5   6   7   8   9  10
  1   9  31 125 206 254 206 108  48  11   1
> hist(num)
> hist(num, breaks=0:11-.5, prob=TRUE, xlim=c(-1,11), ylim=c(0,.3))
> lines(0:10, dbinom(0:10, size=10, prob=1/2), type="o") # cf. Fig. 8.1
```




```
> par(mfrow=c(3,1))
> hist(rbinom(1000, size=10,  prob=1/2)/10,  xlim=c(0,1))
> hist(rbinom(1000, size=30,  prob=1/2)/30,  xlim=c(0,1))
> hist(rbinom(1000, size=100, prob=1/2)/100, xlim=c(0,1))
```

◆ **大数の法則**：（テキスト，p.157 の式(8.5) および p.160 の下段）

任意の大きさの ε に対して，十分に大きな n をとれば，

平均値 $\overline{X}_n \left(= n^{-1} \sum_{i=1\sim n} X_i\right)$ は，以下の関係を満足する．

$$\mathrm{Prob.}\left(\left| \overline{X}_n - \mu \right| < \varepsilon \right) \to 1$$

◎ 大数の法則は，以下の事項に基づく．

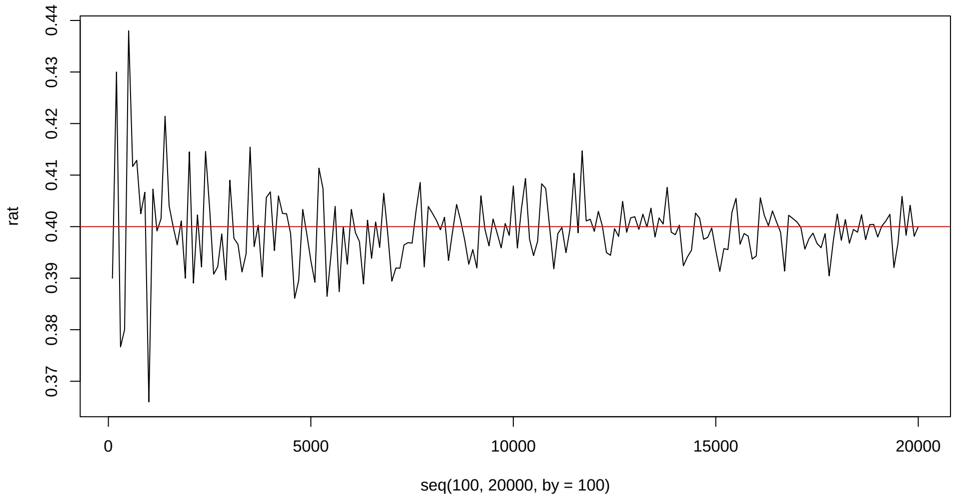
<u>ある分布</u>（平均 $=\mu$, 分散 $=\sigma^2$）から抽出された標本（サイズ n）の標本平均 $\overline{X}$ について，以下が成立する．

$$E\left(\overline{X}_n\right) = \mu; \quad V\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$$

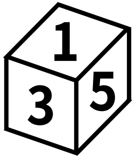
◎ 大数の法則は，さらに精密化できる（中心極限定理）．

試行回数の増加とともに，母比率に近付く様子（テキスト p.158）

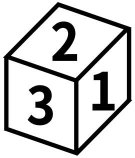
```
> par(mfrow=c(1,1))
> rat <- numeric(0)
> for (trial in seq(100, 20000, by=100)) {
+      rat <- c(rat, rbinom(1, size=trial, prob=0.4)/trial)}
> plot(seq(100, 20000, by=100), rat, type="l")
> abline(h=0.4, col="red") # cf. Fig. 8.2 ~ 8.5
```

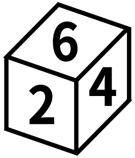


次に，「サイコロの目の和」について，考える．

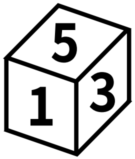
1コ： 

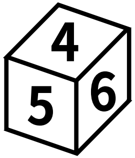
= 1

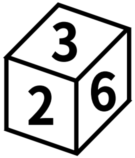
2コ： 

+ 

= 8

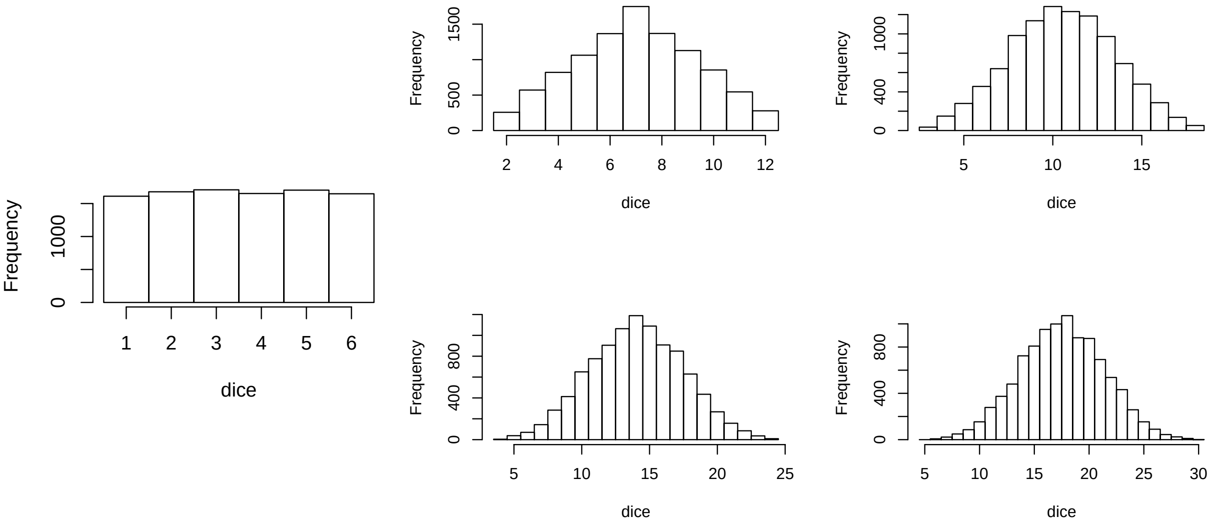
3コ： 

+ 

+ 

= 12

```
> par(mfrow=c(2,2))
> hist(dice <- dice + sample(1:6, 10000, replace=T), breaks=1:12+.5)
> hist(dice <- dice + sample(1:6, 10000, replace=T), breaks=2:18+.5)
> hist(dice <- dice + sample(1:6, 10000, replace=T), breaks=3:24+.5)
> hist(dice <- dice + sample(1:6, 10000, replace=T), breaks=4:30+.5)
> # cf. Fig. 8.7 ~ 8.10
```



◆ **中心極限定理**：（テキスト，p.162 の式(8.6) および p.164, 165 も見よ）

総和 $S_n\left(=\sum_{i=1\sim n} X_i\right)$ に対して，　　$S_n \sim N\left(n\mu, n\sigma^2\right)$

平均 $\overline{X}_n\left(=S_n/n=n^{-1}\sum_{i=1\sim n} X_i\right)$ に対して，　　$\overline{X}_n \sim N\left(\mu, \sigma^2/n\right)$

として，**正規分布**に従うことである．

◎ 中心極限定理は，略して，CLT（Central Limit Theorem） と呼ばれる．

◎ 中心極限定理を大数の法則と対比して，表現すれば，

$$\mathrm{Prob.}\left(\left|\overline{X}_n-\mu\right| < 1.96\,\sigma \quad \sqrt{n}\right) = 0.95$$

となる．つまり，大数の法則では，“記号で表現”されていた箇所が，“値として具体的に”示されている（それは，精密化されたから）．

cf. 大数の法則（再表示）：

$$\mathrm{Prob.}\left(\left|\overline{X}_n-\mu\right| < \varepsilon\right) \to 1$$

◎ 中心極限定理の式表現は，サマザマに可能である．

1） $$\mathrm{Prob.}\left(\left| \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \right| < 1.96 \right) = 0.95$$

2） $$\mathrm{Prob.}\left(\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} < -1.96 \quad or \quad 1.96 < \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \right) = 0.05$$

3） $$\mathrm{Prob.}\left(2.33 < \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \right) = 0.01$$

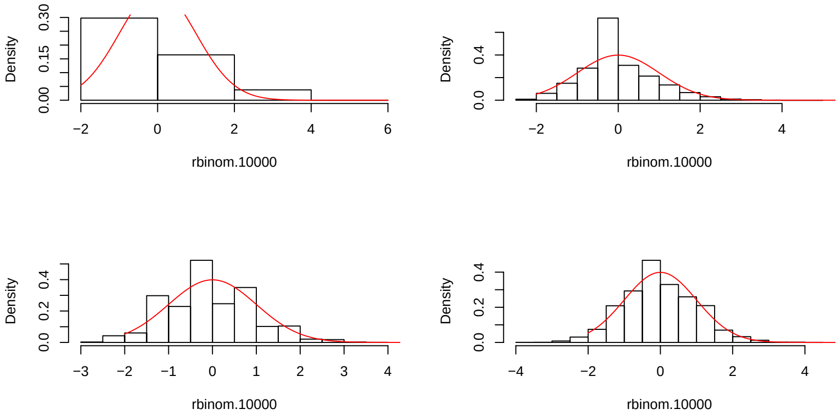
4） $$\mathrm{Prob.}\left(a < \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} < b \right) \to \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-x^2 \;\; 2 \right) dx$$

・・・

つまり，統計量のとりうる範囲を指定すれば，正規分布により，
その確率を算定できる．

中心極限定理の応用（1）：２項分布の正規近似（コインの問題，再び）

```
> hist((rbinom(10000, size=  5, prob=0.1) -   5*0.1)/sqrt(  5*0.1*(1-0.1)))
> hist((rbinom(10000, size= 50, prob=0.1) -  50*0.1)/sqrt( 50*0.1*(1-0.1)))
> hist((rbinom(10000, size=100, prob=0.1) - 100*0.1)/sqrt(100*0.1*(1-0.1)))
> hist((rbinom(10000, size=500, prob=0.1) - 500*0.1)/sqrt(500*0.1*(1-0.1)))
> # Fig. 8.11 ~ 8.16
>
```




```
> pbinom(60, size=500, prob=0.1)
[1] 0.9381745
> pnorm(60, mean=50, sd=sqrt(500*0.1*(1-0.1)))
[1] 0.9319814
```

注：図の縦軸や図中の正規分布などを描画するには，追加コマンドが必要．

２項分布の正規分布による近似：

　２項分布に従う標本 X_i の総和 $S_n = X_1 + X_2 + \ldots + X_i + \ldots + X_n$ について，その平均および分散は，$E(S_n) = np$；$V(S_n) = np(1-p)$ である（成功率 $= p$ ）．例えば，成功率 0.5 として，100 回の試行による成功回数が 40 回以上 60 回以下となる確率は，厳密には，

```
> pbinom(60,    size=100, prob=.5) -
+ pbinom(40 -1, size=100, prob=.5)
[1] 0.9647998
```

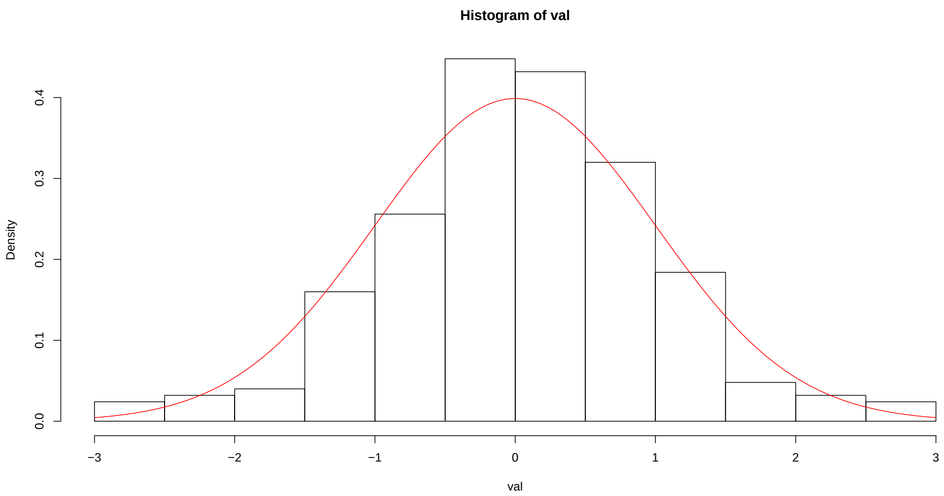
となるのに対し，以下のように，正規分布を用いて近似できるわけである．

```
> pnorm(60, m=.5*100, sd=sqrt(.5*(1-.5)*100)) -
+ pnorm(40, m=.5*100, sd=sqrt(.5*(1-.5)*100))
[1] 0.9544997
```

◆ 注意！！　成功率 p （場合によっては，失敗率）が 0.5 付近で近似の度合いは良好．しかし，成功率 p が，ゼロあるいは１付近では，近似の度合いは悪し！この場合には，２項分布はポアソン分布により近似する（小数の法則；テキスト p.114 を見よ）．

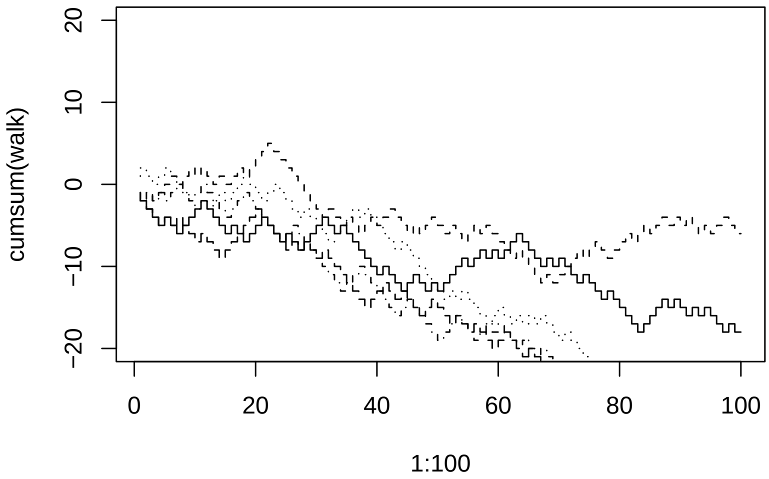
中心極限定理の応用（２）：一様乱数から正規乱数を作る

```
> par(mfrow=c(1,1))
> val <- rep(-6, 250)
> for (h in 1:12) val <- val + runif(250)
> hist(val, prob=TRUE)
> curve(dnorm, -3,3, col="red", add=TRUE)
>
```



ランダム・ウオーク（酔歩）：テキストの練習問題 8.2

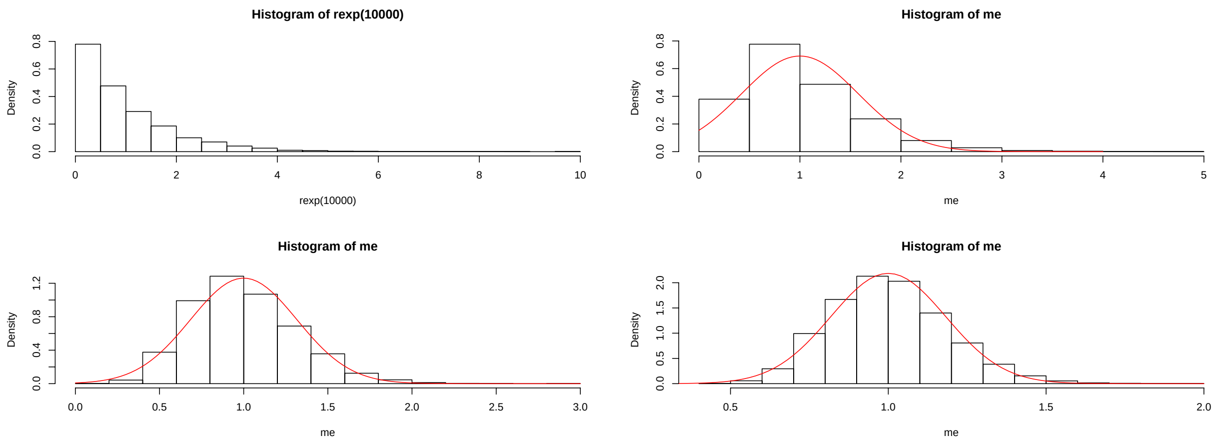
```
> walk <- (rbinom(100, size=1, prob=0.4) - 1/2)*2
> walk[1:10]
 [1]  1  1 -1  1 -1 -1  1  1  1  1
> plot(1:100, cumsum(walk), type="S", ylim=c(-20,20))
> lines(1:100, cumsum((rbinom(100, size=1, prob=0.4) - 1/2)*2), type="S", lty=2)
> ...
```




```
> # for advanced use
> rwalk <- function() lines(1:100, cumsum((rbinom(100, size=1, prob=0.4) - 1/2)*2), type="S")
> rwalk()
```

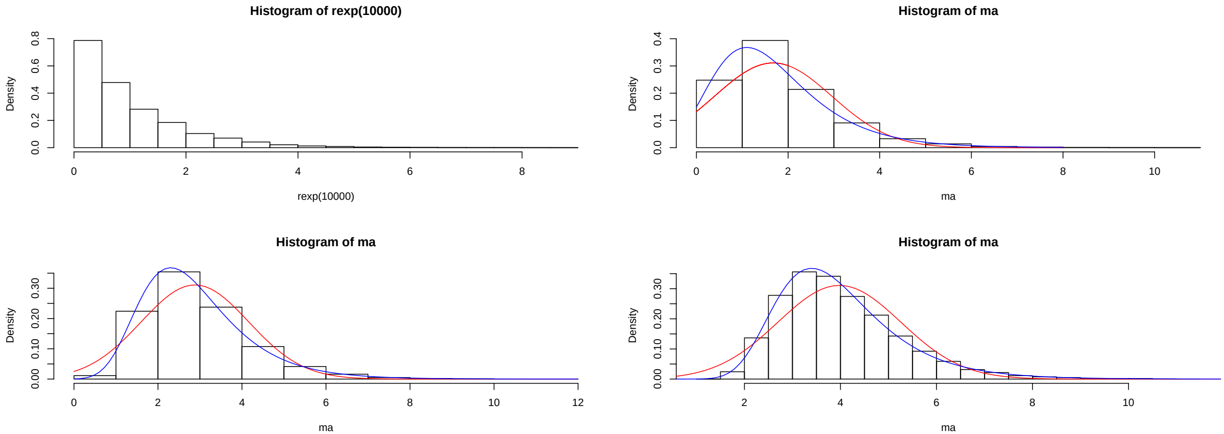
余談：総和あるいは平均という操作に，中心極限定理（CLT）が適用される

```
> hist(rexp(10000), prob=TRUE)
> me <- numeric(10000)
> for (j in 1:10000) me[j] <- mean(rexp(3))
> hist(me, prob=TRUE)
> for (j in 1:10000) me[j] <- mean(rexp(10))
> hist(me, prob=TRUE)
> for (j in 1:10000) me[j] <- mean(rexp(30))
> hist(me, prob=TRUE)
```



余談（続き）：最大値という操作には，CLT が適用されない．別の定理が成立する．

```
> hist(rexp(10000), prob=TRUE)
> ma <- numeric(10000)
> for (j in 1:10000) ma[j] <- max(rexp(3))
> hist(ma, prob=TRUE)
> for (j in 1:10000) ma[j] <- max(rexp(10))
> hist(ma, prob=TRUE)
> for (j in 1:10000) ma[j] <- max(rexp(30))
> hist(ma, prob=TRUE)
```



注：正規分布として近似した赤線は，シミュレーション結果と異なる（青線は極値分布）．