中心極限定理と信頼区間の算出

Table of contents

中心極限定理

中心極限定理

- 平均 μ, 分散 σ^2 の任意の分布から大きさ n の標本を抽出
- 母集団分布の形とは無関係に、標本サイズ n が大きくなるにつれて、標本平均 $\bar{X}_n$ の分布は平均 μ, 分散 σ^2/n 正規分布に近づく
- 母集団が正規分布でなくても「標本平均は正規分布に“収束”する」

Theorem 1 (中心極限定理)

$X_1, X_2, \ldots, X_n$ を独立同分布に従う確率変数列とし、$\mathrm{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$ とする。標本平均 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ について、次が成り立つ：

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \mathrm{N}(0,1) \quad (n \to \infty),$$

すなわち

$$\bar{X}_n \xrightarrow{d} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

中心極限定理のデモンストレーション: 正規分布

演習:

1. 平均 10, 標準偏差 2 の正規分布から大きさ 10 の標本を抽出し、標本平均を算出しましょう
2. 平均 μ, 標準偏差 σ の正規分布から大きさ n の標本を抽出し、標本平均を算出する自作関数を定義しましょう
3. 標本平均の分布を調べるため、平均 10, 標準偏差 2 の正規分布から大きさ 10 の標本を抽出したときの標本平均を 10000 回算出しましょう
4. 生成した標本平均の分布をヒストグラムと正規 Q-Q プロットで確認しましょう
5. 生成した標本平均の平均と分散を計算して中心極限定理を確認しましょう
6. 標本サイズを $5, 10, 50, 200$ と変えた時に標本平均の分布がどの様に変わるか確認しましょう

標本平均

1. 平均 10, 標準偏差 2 の正規分布から大きさ 10 の標本を抽出し、標本平均を算出しましょう

```
mu <- 10
sd <- 2
```
⟶ 平均、標準偏差指定

```
n_sample <- 10
```
⟶ サンプルサイズ（標本の大きさ）指定

```
set.seed(123)
x <- rnorm(n_sample,mean=mu,sd=sd)
```
シード値を固定して標本抽出

```
x_bar <- mean(x)
```
⟶ 標本平均の算出

```
print(x_bar)
```

出力

```
[1] 10.14925
```

自作関数

- R では自由に関数を定義できる

```
関数名 <- function(引数) {
    処理（出力オブジェクトの作成）
    return(出力オブジェクト名)
}
```

  - 引数にデフォルト値を設定できる
    （例：func <- function(x = 1) { ... }）
  - 出力オブジェクトをリストにすることで複数の値を返すことができる
  - 関数内で作った変数は、その関数の中だけでしか使えない
    （以下の例の場合 y は関数外で参照できない）

```
func <- function(x){
    y <- x + 1
    return(y)
}
```

標本平均算出関数

2 平均 μ, 標準偏差 σ の正規分布から大きさ n の標本を抽出し、標本平均を算出する自作関数を定義しましょう

- 入力: 平均 mu, 標準偏差 sigma, 標本の大きさ n
- 出力: 標本平均

```
rnorm_mean <- function(mu,sigma,n){
    x <- rnorm(n,mean=mu,sd=sigma)
    x_bar <- mean(x)
   return(x_bar)
}
```

標本抽出

標本平均の算出

1 で行った計算を関数化

```
class(rnorm_mean)
```

出力

```
[1] "function"
```

標本平均の繰り返し生成

3 平均 10, 標準偏差 2 の正規分布から大きさ 10 の標本を抽出したときの標本平均を 10000 回算出しましょう

```
mu <- 10
sd <- 2
n_sample <- 10
n_sim <- 10000
```
⟶ 標本平均を何回計算するか

```
set.seed(123)
x_bar_vec <- rep(NA,n_sim)
```
⟶ 10000 個の計算した標本平均を入れる空のベクトル

```
for(i in 1:n_sim){
  x_bar_vec[i] <- rnorm_mean(mu=mu,sigma=sd,n=n_sample)
```
⟶ i 番目の要素に計算した標本平均を入れる

```
}
head(x_bar_vec)
tail(x_bar_vec)
```

標本平均の繰り返し生成

出力

```
> head(x_bar_vec)
[1] 10.149251 10.417244 9.150882 10.644089 9.982569
    10.443372
> tail(x_bar_vec)
[1] 10.199367 10.340047 9.264017 10.873002 9.902884
    11.427026
```

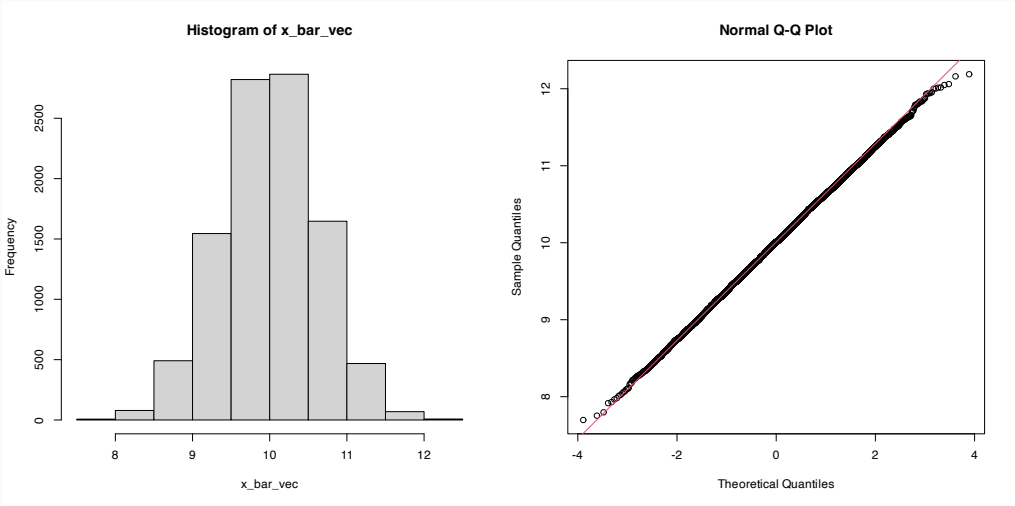
生成した標本平均の視覚化

4 生成した標本平均の分布をヒストグラムと正規 Q-Q プロットで確認しましょう

```
par(mfrow=c(1,2))
```
⟶ 横並びの設定

```
hist(x_bar_vec)
```
⟶ 1 つ目の図：ヒストグラム

```
qqnorm(x_bar_vec)
qqline(x_bar_vec, col=2)
```
2 つ目の図：正規 Q-Q プロット



標本平均の平均と分散

5 生成した標本平均の平均と分散を計算して中心極限定理を確認しましょう

```
mean(x_bar_vec)
mu ⟶ 中心極限定理の主張する標本平均の平均の理論値
var(x_bar_vec)
sd^2/n_sample ⟶ 中心極限定理の主張する標本平均の分散の理論値
```

出力

```
> mean(x_bar_vec)
[1] 10.00195
> mu
[1] 10
> var(x_bar_vec)
[1] 0.394157
> sd^2/n_sample
[1] 0.4
```

中心極限定理のデモンストレーション

6 標本サイズを $5, 10, 50, 200$ と変えた時に標本平均の分布がどの様に変わるか確認しましょう

```
n_sample_vec = c(5,10,50,200)
x_bar_mat = matrix(NA, nrow=n_sim, ncol=length(n_sample
    _vec))
for(j in 1:length(n_sample_vec)){
   n_sample_j = n_sample_vec[j]
   x_bar_vec <- rep(NA,n_sim)
   for(i in 1:n_sim){
      x_bar_vec[i] <- rnorm_mean(mu=mu,sigma=sd,n=n_
          sample_j)
   }
   x_bar_mat[,j] <- x_bar_vec
}
```

中心極限定理のデモンストレーション

```
n_sample_vec = c(5,10,50,200)
```
→ 標本サイズのパターンベクトル

```
x_bar_mat = matrix(NA,
          nrow=n_sim, ncol=length(n_sample_vec))
```
⟶ 各標本サイズのパターンに対して 10000 個の計算した標本平均を入れる空のマトリックス (10000 行 4 列)

```
for(j in 1:length(n_sample_vec)){
  n_sample_j = n_sample_vec[j]
```
⟶ j 番目の標本サイズで固定

```
   x_bar_vec <- rep(NA,n_sim)
   for(i in 1:n_sim){
      x_bar_vec[i] <-
          rnorm_mean(mu=mu,sigma=sd,n=n_sample_j)
    }
```
3 で やった計算

```
  x_bar_mat[,j] <- x_bar_vec
```
⟶ j 列目に 10000 個の標本平均のベクトルを格納

```
}
```

中心極限定理のデモンストレーション

```
title = paste0("Sample size: ", n_sample_vec)
```

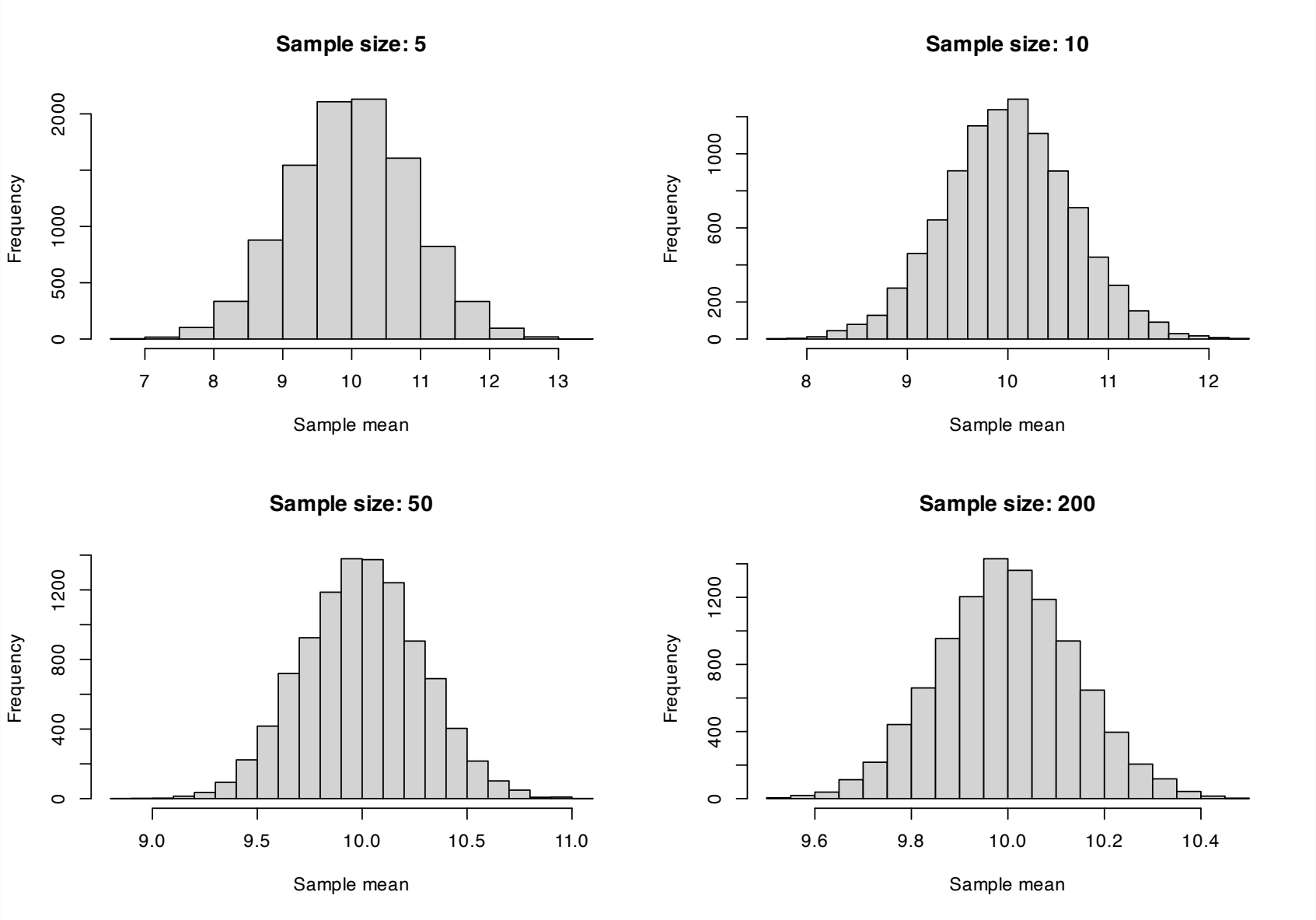
→ 図のタイトルのベクトル

```
par(mfrow=c(2,2))
```
→ 2×2 に並べる

```
for(j in 1:length(n_sample_vec)){
   hist(x_bar_mat[,j],breaks=20,
      main=title[j],xlab="Sample mean")
}
```

→ 標本サイズのパターンごとにヒストグラムを描画

中心極限定理のデモンストレーション



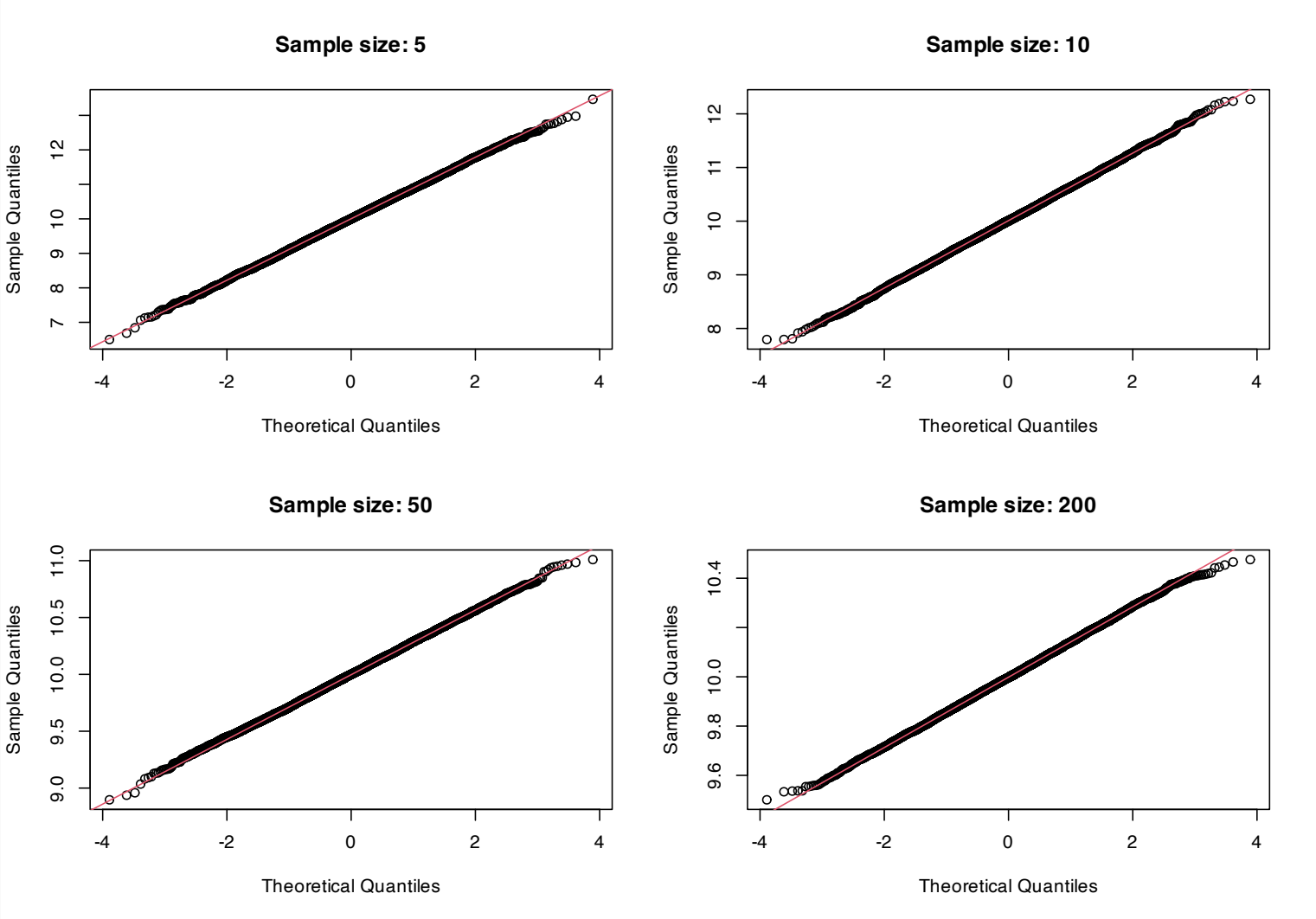
中心極限定理のデモンストレーション

```
par(mfrow=c(2,2))
```
→ 2×2 に並べる

```
for(j in 1:length(n_sample_vec)){
   qqnorm(x_bar_mat[,j],main=title[j])
   qqline(x_bar_mat[,j] ,col=2)
}
```

→ 標本サイズのパターンごとに正規 Q-Q プロットを描画

中心極限定理のデモンストレーション



標本平均の分布は中心極限定理の主張通り、正規分布になっていそう

中心極限定理のデモンストレーション

```
apply(x_bar_mat,2,mean)
mu
apply(x_bar_mat,2,var)
sd^2/n_sample_vec
```

出力

```
> apply(x_bar_mat,2,mean)
[1] 9.996513 10.007144 9.999877 9.998066
> mu
[1] 10
> apply(x_bar_mat,2,var)
[1] 0.78897368 0.39674601 0.07978422 0.02012920
> sd^2/n_sample_vec
[1] 0.80 0.40 0.08 0.02
```

標本平均の分布の平均、分散は中心極限定理の主張とほぼ一致

中心極限定理のデモンストレーション: 様々な分布

演習:

- 以下の分布から標本を抽出するときの標本平均の分布を、標本サイズを $5, 10, 50, 200$ と変えて確認しましょう
  1. 区間 $[0, 1]$ の一様分布 $(\mu = 1/2, \sigma^2 = 1/12)$
  2. パラメータ $\lambda = 2$ の指数分布 $(\mu = 1/\lambda, \sigma^2 = 1/\lambda^2)$
  3. $p = 0.3$ のベルヌーイ分布 $(\mu = p, \sigma^2 = p(1-p))$
  4. パラメータ $\lambda = 3$ のポアソン分布 $(\mu = \lambda, \sigma^2 = \lambda)$

  手順

  - 標本平均を 10000 回生成
  - ヒストグラムと正規 Q-Q プロットで視覚化
  - 平均と分散を計算

  (p.12 のコードの関数 `rnorm_mean` の部分のみ対応する分布のものに変えれば OK)

コード例: 一様分布

・関数定義

```
uniform_mean <- function(min=0,max=1,n){
  x <- runif(n,min=min,max=max)
  x_bar <- mean(x)
  return(x_bar)
}

min = 0
max = 1
mu = (max + min)/2
var = (max-min)^2/12
```

コード例: 一様分布

・乱数生成: 自作関数の箇所を変えるだけ

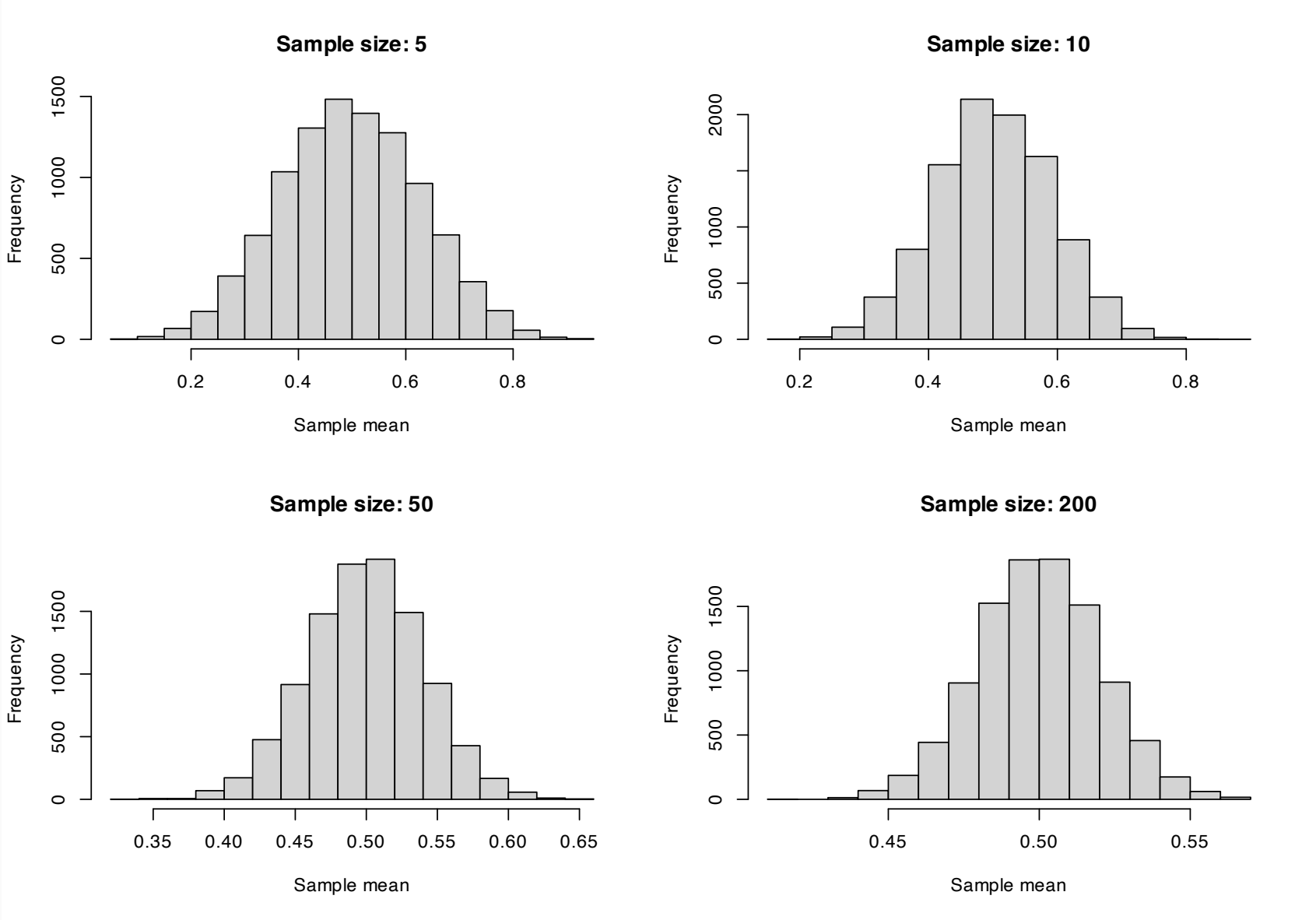
```
set.seed(123)
n_sample_vec = c(5,10,50,200)
x_bar_mat = matrix(NA, nrow=n_sim, ncol=length(n_sample
    _vec))
for(j in 1:length(n_sample_vec)){
  n_sample_j = n_sample_vec[j]
  x_bar_vec <- rep(NA,n_sim)
  for(i in 1:n_sim){
    x_bar_vec[i] <- uniform_mean(n=n_sample_j)
  }
  x_bar_mat[,j] = x_bar_vec
}
```

コード例: 一様分布

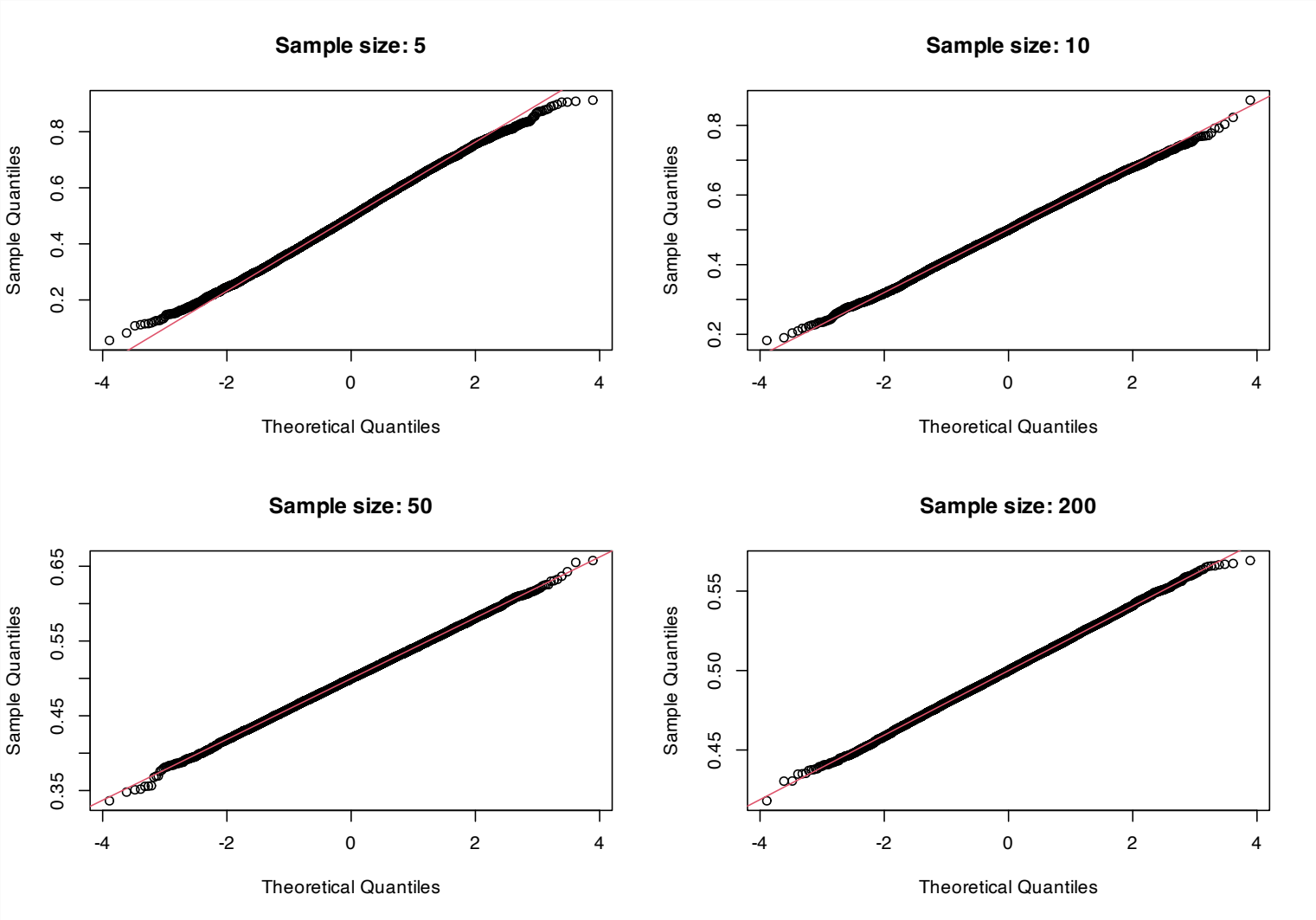
- 視覚化: 同じコード

```
title = paste0("Sample size: ", n_sample_vec)
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   hist(x_bar_mat[,j],breaks=20,
   main=title[j],xlab="Sample mean")
}
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   qqnorm(x_bar_mat[,j],main=title[j])
   qqline(x_bar_mat[,j] ,col=2)
}
```

出力例ヒストグラム: 一様分布



出力例 Q-Q plot: 一様分布



サンプルサイズが小さいとき、少しずれていそう

出力例平均分散: 一様分布

```
apply(x_bar_mat,2,mean)
mu
apply(x_bar_mat,2,var)
var/n_sample_vec
```

出力

```
> apply(x_bar_mat,2,mean)
[1] 0.4973980 0.5010886 0.4995747 0.5000017
> mu
[1] 0.5
> apply(x_bar_mat,2,var)
[1] 0.0167238793 0.0081600900 0.0016493200 0.0004191163
> var/n_sample_vec
[1] 0.0166666667 0.0083333333 0.0016666667 0.0004166667
```

標本平均の分布の平均、分散は中心極限定理の主張とほぼ一致

コード例: 指数分布

- 関数定義

```
exp_mean <- function(lambda,n){
  x <- rexp(n,rate=lambda)
  x_bar <- mean(x)
  return(x_bar)
}
lambda = 2
mu = 1 / lambda
var = 1 / lambda^2
```

コード例: 指数分布

・乱数生成: 自作関数の箇所を変えるだけ

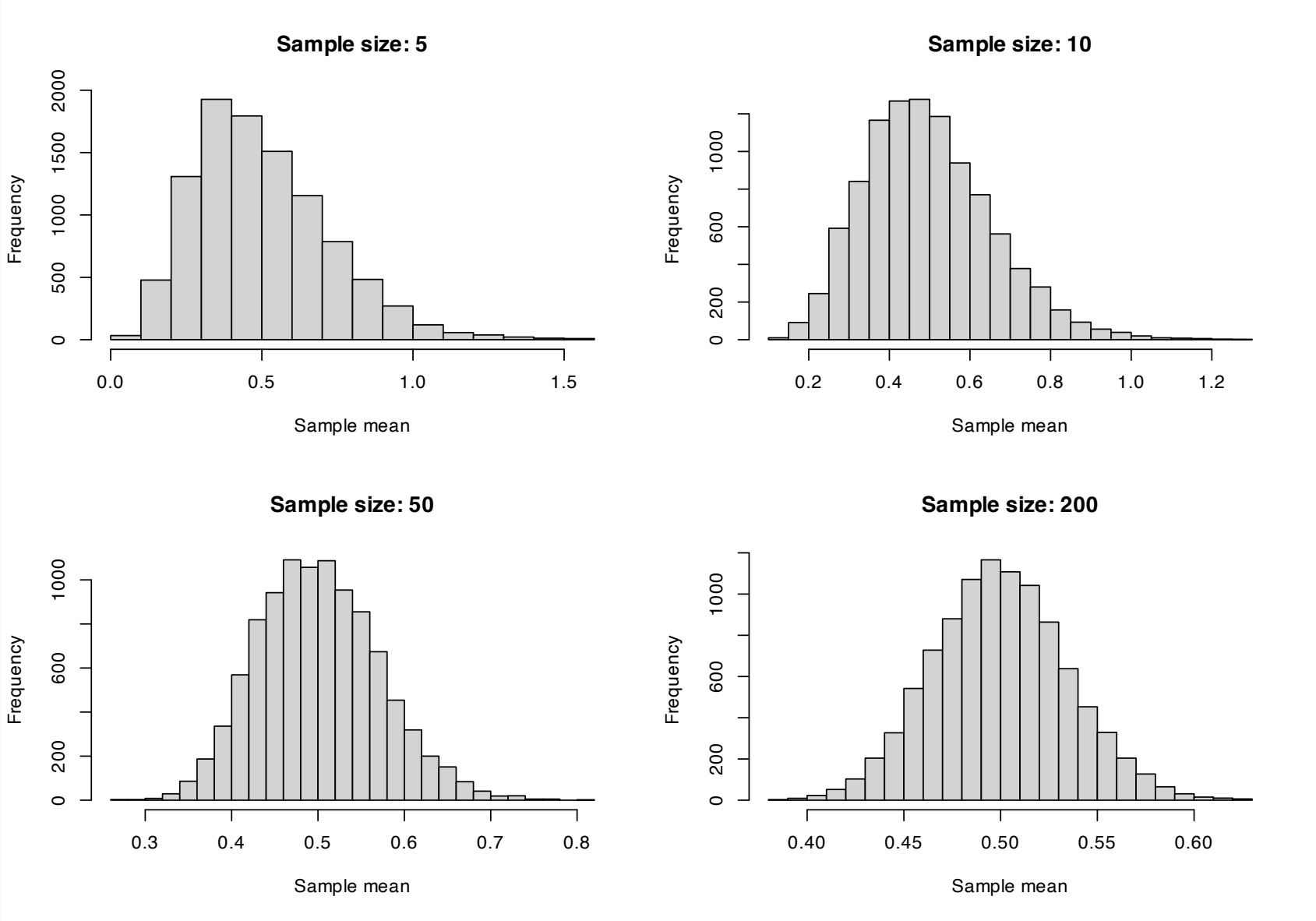
```
set.seed(123)
n_sample_vec = c(5,10,50,200)
x_bar_mat = matrix(NA, nrow=n_sim, ncol=length(n_sample
    _vec))
for(j in 1:length(n_sample_vec)){
  n_sample_j = n_sample_vec[j]
  x_bar_vec <- rep(NA,n_sim)
  for(i in 1:n_sim){
    x_bar_vec[i] <-
        exp_mean(lambda=lambda,n=n_sample_j)
  }
  x_bar_mat[,j] = x_bar_vec
}
```

コード例: 指数分布

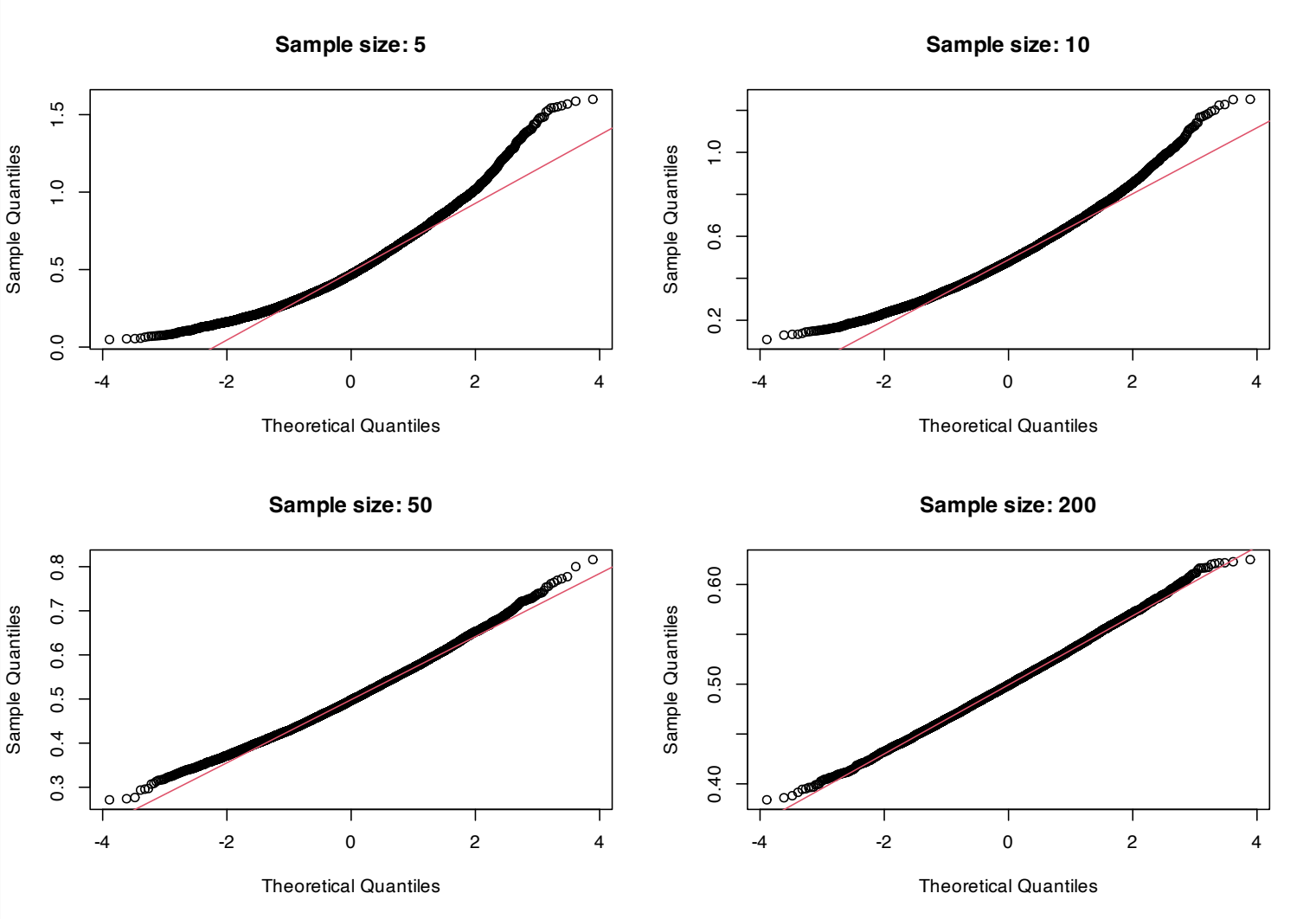
- 視覚化: 同じコード

```
title = paste0("Sample size: ", n_sample_vec)
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   hist(x_bar_mat[,j],breaks=20,
   main=title[j],xlab="Sample mean")
}
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   qqnorm(x_bar_mat[,j],main=title[j])
   qqline(x_bar_mat[,j] ,col=2)
}
```

出力例ヒストグラム: 指数分布



出力例 Q-Q plot: 指数分布



サンプルサイズ 200 でほぼ正規分布

出力例平均分散: 指数分布

```
apply(x_bar_mat,2,mean)
mu
apply(x_bar_mat,2,var)
var/n_sample_vec
```

出力

```
> apply(x_bar_mat,2,mean)
[1] 0.5016903 0.4961924 0.5006024 0.4997620
> mu
[1] 0.5
> apply(x_bar_mat,2,var)
[1] 0.049696731 0.024999329 0.004966961 0.001211390
> var/n_sample_vec
[1] 0.05000 0.02500 0.00500 0.00125
```

標本平均の分布の平均、分散は中心極限定理の主張とほぼ一致

コード例: ベルヌーイ分布

・関数定義

```
bernoulli_mean <- function(p,n){
  x <- rbinom(n,size=1, prob=p)
  x_bar <- mean(x)
  return(x_bar)
}
p = 0.3
mu = p
var = (1-p)*p
```

コード例: ベルヌーイ分布

- 乱数生成: 自作関数の箇所を変えるだけ

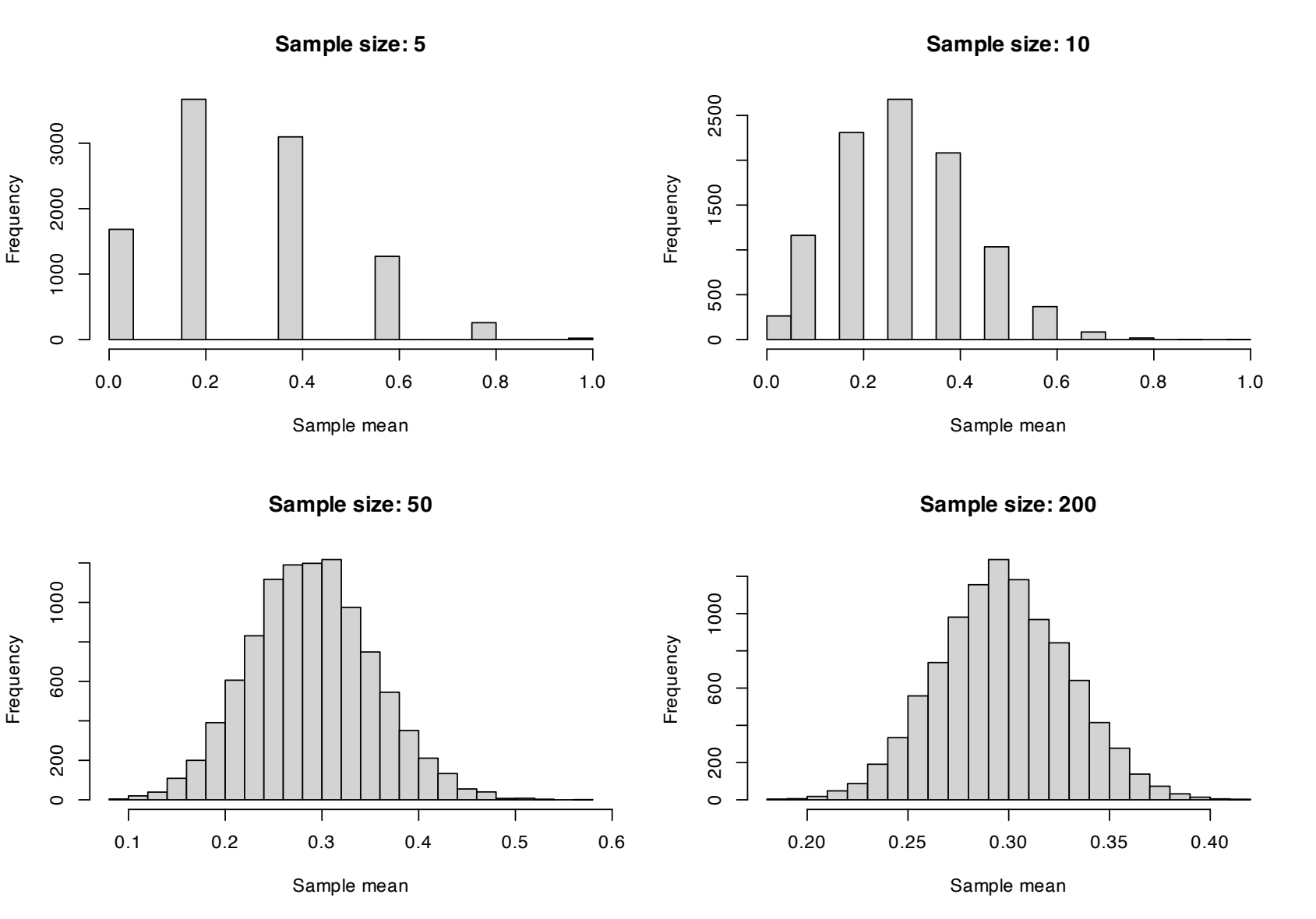
```
set.seed(123)
n_sample_vec = c(5,10,50,200)
x_bar_mat = matrix(NA, nrow=n_sim, ncol=length(n_sample
    _vec))
for(j in 1:length(n_sample_vec)){
  n_sample_j = n_sample_vec[j]
  x_bar_vec <- rep(NA,n_sim)
  for(i in 1:n_sim){
    x_bar_vec[i] <- bernoulli_mean(p=p,n=n_sample_j)
  }
  x_bar_mat[,j] = x_bar_vec
}
```

コード例: ベルヌーイ分布

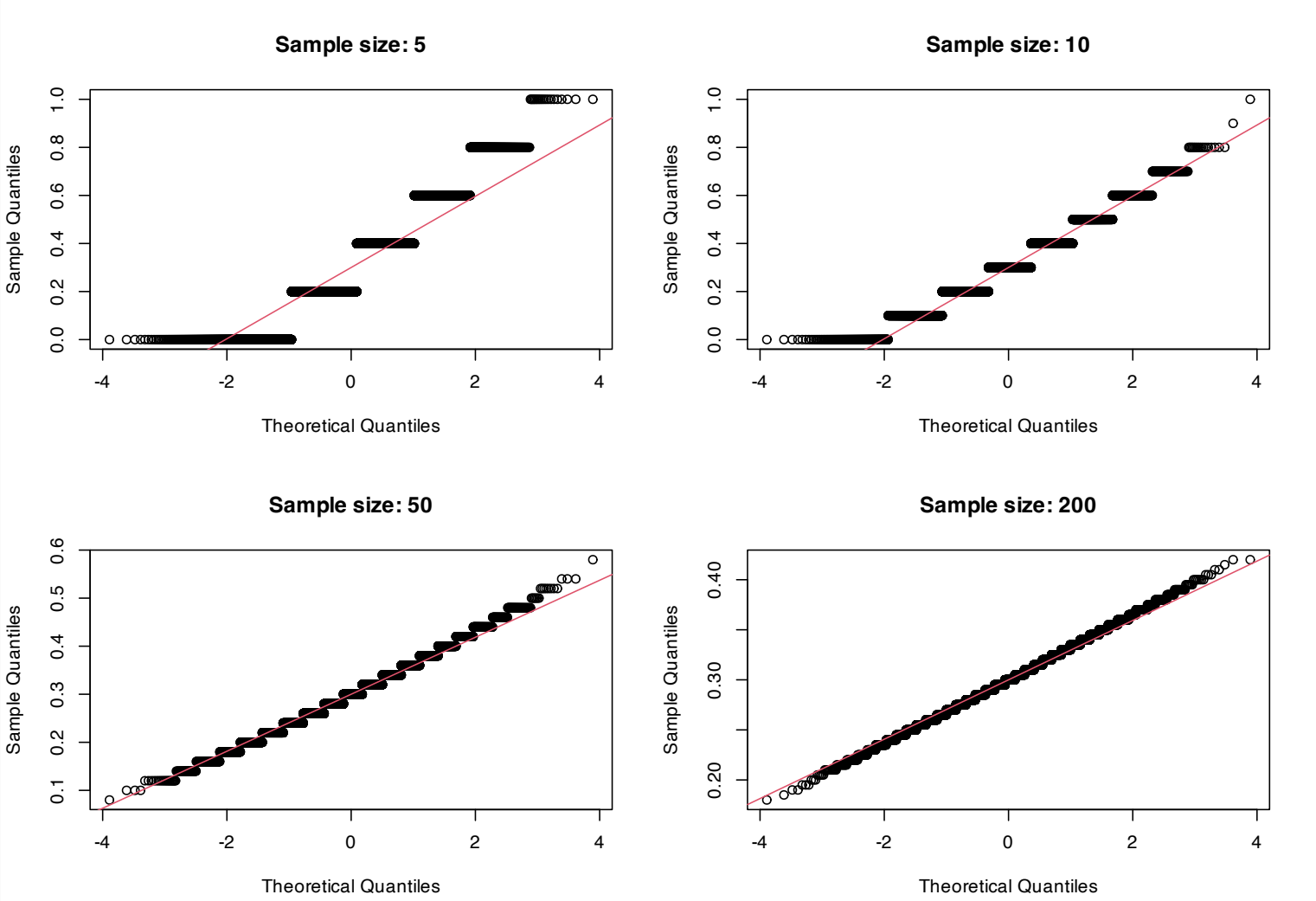
- 視覚化: 同じコード

```
title = paste0("Sample size: ", n_sample_vec)
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   hist(x_bar_mat[,j], breaks=20,
   main=title[j], xlab="Sample mean")
}
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   qqnorm(x_bar_mat[,j], main=title[j])
   qqline(x_bar_mat[,j], col=2)
}
```

出力例ヒストグラム: ベルヌーイ分布



出力例 Q-Q plot: ベルヌーイ分布



離散的、サンプルサイズ 200 でほぼ正規分布

出力例平均分散: ベルヌーイ分布

```
apply(x_bar_mat,2,mean)
mu
apply(x_bar_mat,2,var)
var/n_sample_vec
```

出力

```
> apply(x_bar_mat,2,mean)
[1] 0.2961200 0.3026300 0.2995240 0.2998925
> mu
[1] 0.3
> apply(x_bar_mat,2,var)
[1] 0.040757021 0.020694153 0.004088502 0.001044625
> var/n_sample_vec
[1] 0.04200 0.02100 0.00420 0.00105
```

標本平均の分布の平均、分散は中心極限定理の主張とほぼ一致

コード例: ポアソン分布

・関数定義

```
poisson_mean <- function(lambda,n){
  x <- rpois(n,lambda=lambda)
  x_bar <- mean(x)
  return(x_bar)
}
lambda = 3
mu = lambda
var = lambda
```

コード例: ポアソン分布

・乱数生成: 自作関数の箇所を変えるだけ

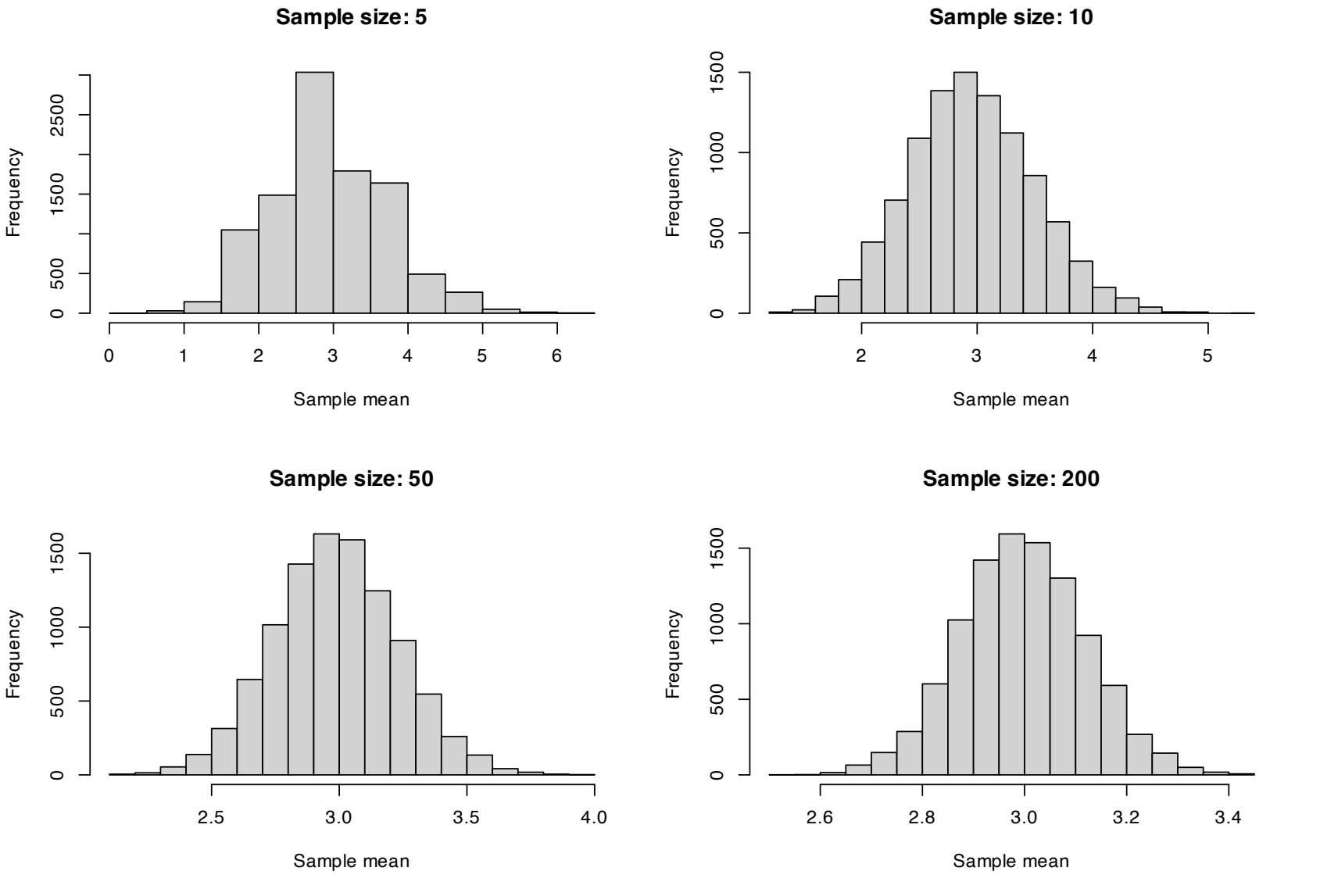
```
set.seed(123)
n_sample_vec = c(5,10,50,200)
x_bar_mat = matrix(NA, nrow=n_sim, ncol=length(n_sample
    _vec))
for(j in 1:length(n_sample_vec)){
  n_sample_j = n_sample_vec[j]
  x_bar_vec <- rep(NA,n_sim)
  for(i in 1:n_sim){
    x_bar_vec[i] <-
        poisson_mean(lambda=lambda,n=n_sample_j)
  }
  x_bar_mat[,j] = x_bar_vec
}
```

コード例: ポアソン分布

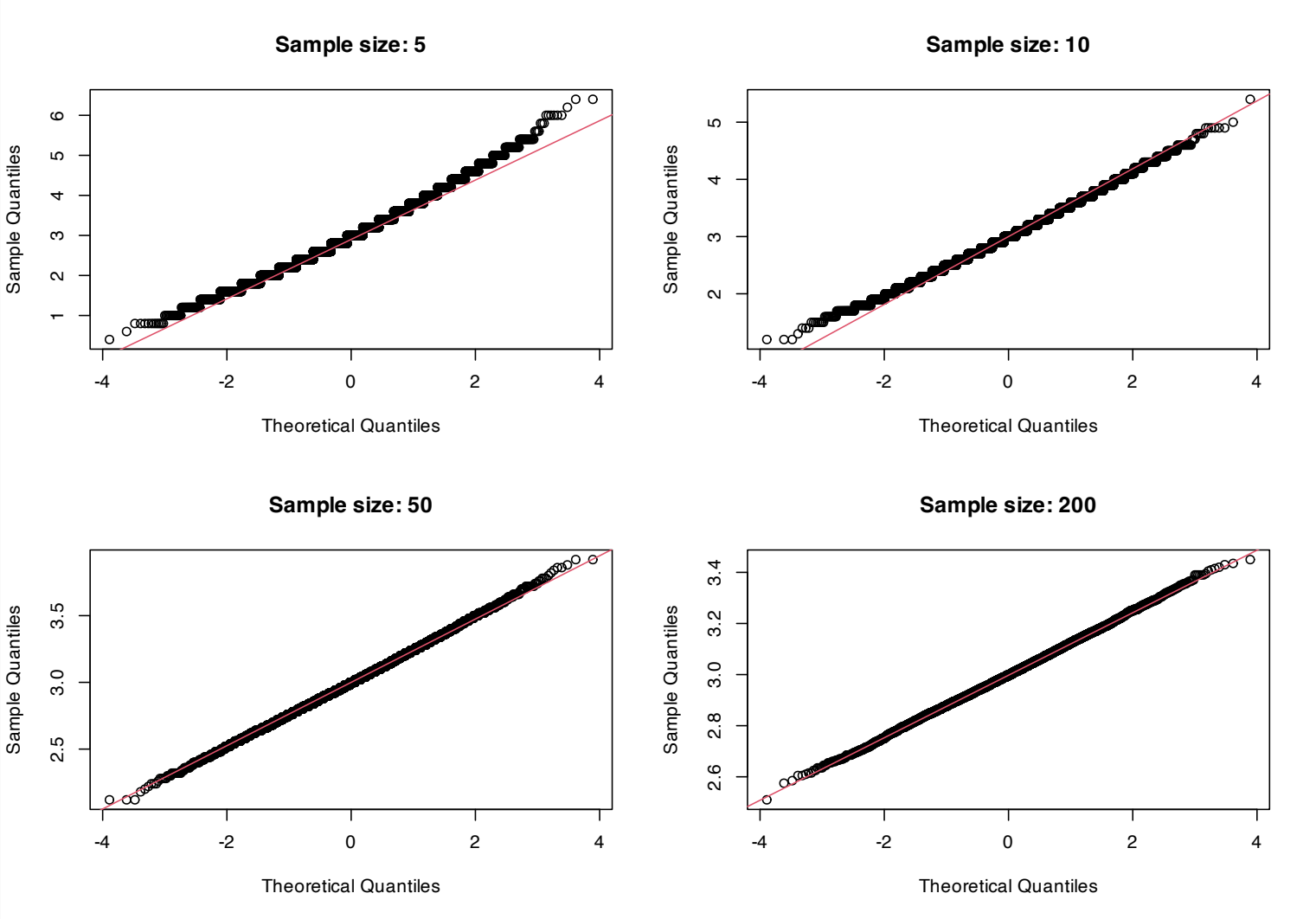
- 視覚化: 同じコード

```
title = paste0("Sample size: ", n_sample_vec)
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   hist(x_bar_mat[,j], breaks=20,
   main=title[j], xlab="Sample mean")
}
par(mfrow=c(2,2))
for(j in 1:length(n_sample_vec)){
   qqnorm(x_bar_mat[,j], main=title[j])
   qqline(x_bar_mat[,j], col=2)
}
```

出力例ヒストグラム: ポアソン分布



出力例 Q-Q plot: ポアソン分布



サンプルサイズ 50 でほぼ正規分布

出力例平均分散: ポアソン分布

```
apply(x_bar_mat,2,mean)
mu
apply(x_bar_mat,2,var)
var/n_sample_vec
```

出力

```
> apply(x_bar_mat,2,mean)
[1] 2.988520 3.005100 2.997460 2.999519
> mu
[1] 3
> apply(x_bar_mat,2,var)
[1] 0.60375258 0.29412540 0.05927180 0.01506729
> var/n_sample_vec
[1] 0.600 0.300 0.060 0.015
```

標本平均の分布の平均、分散は中心極限定理の主張とほぼ一致

信頼区間の算出

リスク差の信頼区間 (中心極限定理の応用)

例: 禁煙治療データ（asaur パッケージ, pharmacoSmoking）

- 治療群（grp=combination）でのイベント（禁煙失敗）確率を p_1, 対照群（grp=patchOnly）でのイベント確率を p_2
- それぞれの群のサンプルサイズを n_1, n_2 とする
- 治療群 $(j = 1)$、対照群 $(j = 2)$ の被験者 $i = 1, \ldots, n_j$ の禁煙失敗かどうかを表す変数を X_{ij} とする
- X_{ij} は確率 p_j のベルヌーイ分布に従うと考えられる $(j = 1, 2)$
- リスク差の推定値は各群の標本平均 $\bar{X}_j$ の差で表され、中心極限定理より以下の正規分布で近似できる

$$\hat{p}_1 - \hat{p}_2 = \frac{\sum_{i=1}^{n_1} X_{i1}}{n_1} - \frac{\sum_{i=1}^{n_2} X_{i2}}{n_2} = \bar{X}_1 - \bar{X}_2 \approx \mathrm{N}\Big(p_1 - p_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\Big).$$

- リスク差 $p_1 - p_2$ の 95% 信頼区間

$$(\bar{X}_1 - \bar{X}_2) \pm z_{0.975} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{1}$$

ここで $\sigma_j^2 = p_j(1 - p_j)$.

2×2 分割表による関連指標と区間推定

R パッケージ `epiR` の関数 `epi.2by2`

- 分割表を入力しリスク差、リスク比、オッズ比等の点推定、区間推定等を出力する関数
- どの様な試験デザインで得られた表かを引数 `method` に指定
- 分割表は以下の形式（順番）が想定されている
  - コホート研究: `method="cohort.count"`[1]
  - 症例対照研究: `method="case.control"`
  - 横断研究 `method="cross.sectional"`

	Disease +	Disease -	Total
Expose +	a	b	$a+b$
Expose -	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

[1]追跡時間付きのデータの場合は `method="cohort.time"` を使用（分割表の形式は manual 参照）

2×2 分割表：信頼区間の算出例

```
install.packages("epiR")
library(epiR)
library(asaur)
data("pharmacoSmoking")

group = pharmacoSmoking$grp
event = factor(pharmacoSmoking$relapse,levels=c(1, 0),
    labels=c("relapse","abstinence"))
tab <- table(group, event)
print(tab)

epi.2by2(tab, method="cohort.count", conf.level=0.95)
```

演習: 上記のコードを実行してリスク差の信頼区間を確認しましょう

出力例

```
> epi.2by2(tab, method="cohort.count", conf.level=0.95)
                Outcome+    Outcome-       Total                  Inc risk *
Exposure+             37          24          61   60.66 (47.31 to 72.93)
Exposure-             52          12          64   81.25 (69.54 to 89.92)
Total                 89          36         125   71.20 (62.42 to 78.95)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 0.75 (0.59, 0.94)
Inc odds ratio                                 0.36 (0.16, 0.80)
Attrib risk in the exposed *                   -20.59 (-36.14, -5.05)
Attrib fraction in the exposed (%)            -33.95 (-72.49, -6.97)
Attrib risk in the population *                -10.05 (-22.48, 2.38)
Attrib fraction in the population (%)         -14.12 (-13.90, -11.40)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 6.460 Pr>chi2 = 0.011
Fisher exact test that OR = 1: Pr>chi2 = 0.017
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Incidence risk:
イベント発生確率

p_1

p_2

分割表

リスク比

オッズ比

リスク差

点推定値
と
信頼区間

検定

連続量の平均差の信頼区間 (R での算出方法)

- `t.test()`: 連続量の 2 群の平均差 $\mu_1 - \mu_2$ の 95% 信頼区間を算出
  - 本来検定のための関数だが信頼区間も計算される
  - デフォルトでは等分散を仮定せずに推定
  - 等分散を仮定する場合は `var.equal = TRUE` を指定
- 返り値の `$conf.int` に信頼区間が格納される

連続量の平均差の信頼区間：コード例

```
set.seed(123)
group1 <- rnorm(20, mean = 5, sd = 2)
group2 <- rnorm(25, mean = 3, sd = 3)

res <- t.test(group1, group2, conf.level = 0.95)
print(res)
print(res$conf.int)
```

演習: 上記のコードを実行して平均の差の信頼区間を確認しましょう

出力例

```
> print(res)

        Welch Two Sample t-test

data:  group1 and group2
t = 3.1643, df = 42.131, p-value = 0.002885
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8192129 3.7031813
sample estimates:
mean of x mean of y
 5.283248  3.022050

> print(res$conf.int)
[1] 0.8192129 3.7031813
attr(,"conf.level")
[1] 0.95
```

検定の結果

区間推定

点推定

相関係数の区間推定

- ピアソンの相関係数: ベース R の `cor.test()` で信頼区間算出
  - `method="pearson"` の指定
  - Fisher の z 変換で正規近似して信頼区間を構成
- スピアマンの相関係数: ライブラリ `spearmanCI()` の `spearmanCI()` で信頼区間算出
  - `cor.test()` での `method="spearman"` の指定は信頼区間は計算しない（検定のみ）
- 帰無仮説 (相関ゼロ) の元での分布と相関を限定しない元での分布を異なる近似を用いて計算

相関係数の信頼区間算出例

```
install.packages("spearmanCI")
library(spearmanCI)
library(mlbench)
data("PimaIndiansDiabetes2")
data <- na.omit(PimaIndiansDiabetes2)
x = data$glucose
y = data$insulin

cor(x, y, method="pearson")
cor.test(x, y, method="pearson")

cor(x, y, method="spearman")
spearmanCI(x,y)
```

演習: 上記のコードを実行して相関係数の信頼区間を確認しましょう

出力例

```
> cor(x, y, method="pearson")
[1] 0.581223
> cor.test(x, y, method="pearson")

        Pearson's product-moment correlation

data:  x and y
t = 14.105, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5116288 0.6432405
sample estimates:
      cor
0.581223

> cor(x, y, method="spearman")
[1] 0.6589582
> spearmanCI(x,y)
confidence interval
     2.5 %    97.5 %
 0.5952308 0.7226585
sample estimate
 0.6589582
```

検定の結果

区間推定

点推定

区間推定

点推定

参考

（参考）リスク差の信頼区間

⑴ 式を用いてリスク差の信頼区間を算出し、epi.2by2 の結果と一致することを確認せよ

```
tb <- table(pharmacoSmoking$grp,
          pharmacoSmoking$relapse)
ptb <- prop.table(tb, margin=1)
p1 <- ptb[1,2]
p2 <- ptb[2,2]

var1 <- p1 * (1-p1)
var2 <- p2 * (1-p2)
n1 <- table(pharmacoSmoking$grp)[1]
n2 <- table(pharmacoSmoking$grp)[2]

(p1 - p2)
(p1 - p2) + qnorm(0.975) * sqrt(var1/n1 + var2/n2)
(p1 - p2) - qnorm(0.975) * sqrt(var1/n1 + var2/n2)
```

（参考）中心極限定理を用いた群間平均差の信頼区間の導出

- 母平均 μ_1, μ_2、母分散 σ_1^2, σ_2^2 の二つの母集団から大きさ n_1, n_2 の独立標本を抽出し、それぞれの標本平均を $\bar{X}_1, \bar{X}_2$ とする。
- 中心極限定理により、それぞれの群で以下が成り立つ

$$\bar{X}_j \approx \mathrm{N}\left(\mu_j, \frac{\sigma_j^2}{n_j}\right) \quad (j = 1, 2).$$

- 独立性から差 $D = \bar{X}_1 - \bar{X}_2$ は以下の正規分布に従う

$$D \approx \mathrm{N}\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$$

- 整理すると以下の統計量が標準正規分布に従う

$$\frac{D - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \approx \mathrm{N}(0, 1)$$

（参考）中心極限定理を用いた群間平均差の信頼区間の導出

- 標準正規分布の上側確率 2.5% となる点を $z_{0.975}$ とすると、

$$\Pr\left(-z_{0.975} \leq \frac{D-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \leq z_{0.975}\right) = 0.95$$

- 内部の不等式を整理すると

$$\Pr\left(D - z_{0.975}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2} \leq \mu_1-\mu_2 \right.$$

$$\left. \leq D + z_{0.975}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}\right) = 0.95$$

- 群間平均差 $\mu_1-\mu_2$ の信頼区間は下記

$$(\bar{X}_1-\bar{X}_2) \pm z_{0.975}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

- σ_1^2, σ_2^2 は未知なので推定値を代入

（参考）連続量の群間平均差の信頼区間

- 近似を用いた信頼区間: 中心極限定理より

$$\frac{D-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \approx \mathrm{N}(0,1)$$

  σ_1^2, σ_2^2 に推定値を代入して信頼区間を算出

$$D \pm z_{0.975}\sqrt{\hat{\sigma}_1^2/n_1+\hat{\sigma}_2^2/n_2}$$

- 正確な分布を用いた信頼区間: σ_1^2, σ_2^2 に推定値を代入した統計量の分布を導出: 自由度 ν の t 分布に従う

$$\frac{D-(\mu_1-\mu_2)}{\sqrt{\hat{\sigma}_1^2/n_1+\hat{\sigma}_2^2/n_2}} \sim T_\nu$$

  自由度 ν の t 分布の上側 2.5% を $t_{0.975}$ として、以下の式から信頼区間を算出

$$\Pr\left(-t_{0.975} \leq \frac{D-(\mu_1-\mu_2)}{\sqrt{\hat{\sigma}_1^2/n_1+\hat{\sigma}_2^2/n_2}} \leq t_{0.975}\right) = 0.95$$