4.3.3　What does the joint density look like?

＊対応のある２つのベクトルがある場合の結合密度はどのようなものか

確率変数が１つの密度の場合：密度曲線と X 軸で囲まれた面積が１に等しい。

対応のある２つのベクトルがある場合：密度は表面で、表面密度と X 軸と Y 軸で囲まれた面の間の容積が１に等しい

Figure 4.14　左上の図:1000 個の２変量の標準正規変量の無作為標本の密度を示している。

＊２変量の正規乱数を生成するための関数

MASS パッケージをロード→関数 mvrnorm( )（２つ以上の相関のあるベクトルを生成する関数)を利用可能にする。

関数 mvrnorm( )を使用して、平均 0、分散 1、相関 0.8 の母集団から標本抽出された n=1000 の対応のある乱数の無作為標本を生成

```
> library(MASS)
> x = mvrnorm(n = 1000, mu = c(0,0), Sigma = cbind(c(1, 0.8), c(0.8, 1)))
> head(x)
```

cor ( ) を使用して、２つの列のベクトルの相関が母集団のパラメータ 0.8 に実際近いかを確認する

```
> cor(x[, 1], x[, 2])
```

３番目の引数 Sigma（ベクトルを列方向に結びつける cbind ( )で生成される）は、２変量の標準正規標本 x の **variance-covariance matrix**（分散共分散行列）である。

```
> Sigma = cbind(c(1, 0.8), c(0.8, 1))
> Sigma
```

Sigma は、主対角線上に分散、副対角線上に共分散

※相関と共分散の違い

２つの列の相関は、ベクトルの規模の増減に関係なく同じであるが、共分散はベクトルの規模の変化により大幅に変化する。

```
> cor(x[, 1], x[, 2])
> cor(x[, 1], 100 * x[, 2])
> cor(0.001 * x[, 1], 100 * x[, 2])
```

```
> cov(x[, 1], x[, 2])
> cov(x[, 1], 100 * x[, 2])
> cov(0.003 * x[, 1], 100 * x[, 2])
```

＊関数 kde2d( )で２変量の正規乱数の表面密度を推定し、pers ( )で透視図を作成する

```
> persp(kde2d(x[, 1], x[, 2], n = 50), phi = 30, theta = 20, d = 10, col ="lightblue",
+ shade = 0.75, ltheta = -100, border = NA, expand = 0.5,
+ xlab = "x", ylab = "Y", zlab = "density")
> mtext("bivariate standard normal", 3, 1)
```

↑訂正（テキストは+)

＊Figure 4. 14 右上の図：

対数正規のポアソン分布(**lognormal-Poisson distributed**)の 2 変量の密度

これは、たとえば 2 つの同じサイズのコーパスにおける 1 組の語の頻度を計算することによって得られる対応のある語の頻度に対する最初の近似値を示す分布

対数正規確率変数(**lognormal random variable**)：対数変換の後正規分布している変量

語の頻度は対数正規分布しているという仮説→rlnorm( )を使用して、n=1000 の対数正規分布した乱数を生成　rlnorm( ): 1000 語がテキストで使われているポアソン比 λ をモデル化

＊2 つのコーパスにおけるある語の頻度をシミュレートするために、rpois ( )を用いてその語に対する 2 つの乱数を生成する

```
> n = 1000
> lambdas =rlnorm(n, 1, 4)
> mat = matrix(nrow = n, ncol =2)
```

※for loop で処理（mat の各単語 i に対する 2 つの頻度を格納する）

```
> for (i in 1:n) { mat[i, ] = rpois(2, lambdas[i])}
> mat[1:10, ]
```

※mat の頻度 0 をなくすために、すべてのセルに 1 足して、対数変換する

```
> mat = log(mat+1)
```

※透視図作成

```
> persp(kde2d(mat[, 1], mat[, 2], n = 50), phi = 30, theta = 20, d = 10, col = "lightblue",
+ shade = 0.75, box = T, border = NA,  ltheta = -100, expand = 0.5,
+ xlab = "log x", ylab = "log Y", zlab = "density")
> mtext("bivariate lognormal-Poisson", 3, 1)
```

＊Figure 4.14 の下の図： 2 つの実証的な密度

左下の図： 4 つの音素を持つ 4171 のオランダ語の単語音韻的類似性に関するもの

1 つの音素のみ異なる（音韻的近似） 4 つの音素を持つ異なり語の数を計算

近似の語のランクを計算（最も頻度が高い→ランクは 1 ）

→近似の語がないものを除き、対数変換したあと密度を得る

右下の図：4633 のオランダ語の単一形態素の名詞の単数形と複数形の（対数）頻度の密度分布は、右上の対数正規のポアソン分布と同じ種類の形をしている。