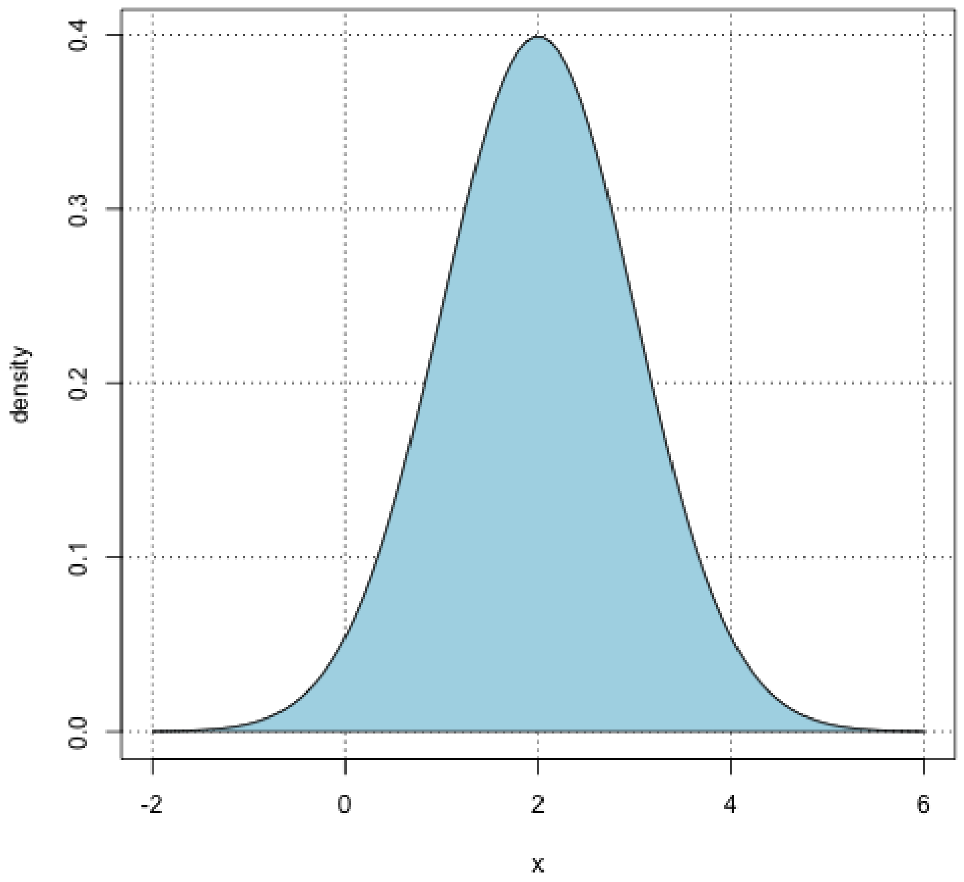
MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Multivariate normal distribution

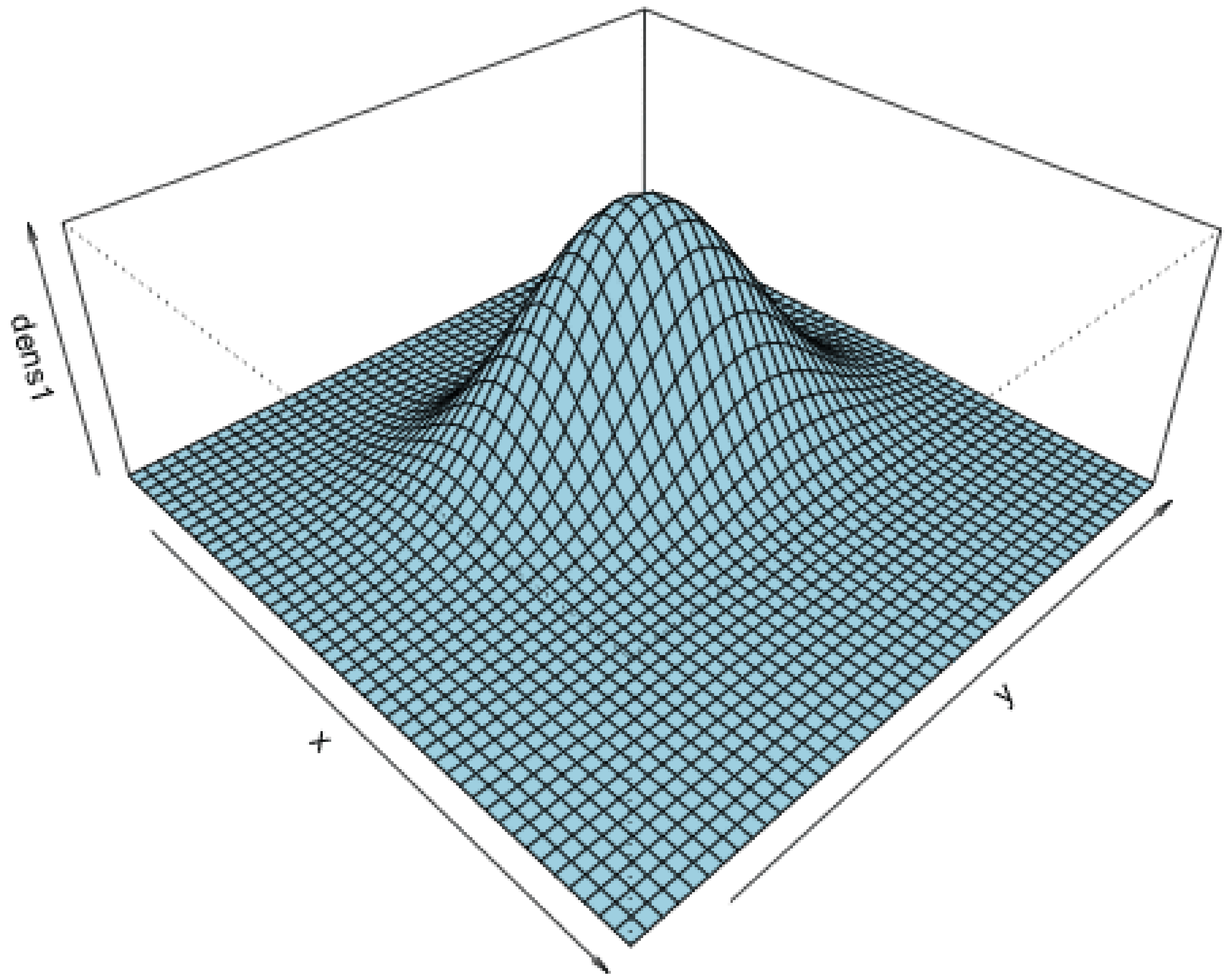
Surajit Ray
Senior Lecturer, University of Glasgow

Univariate normal distribution



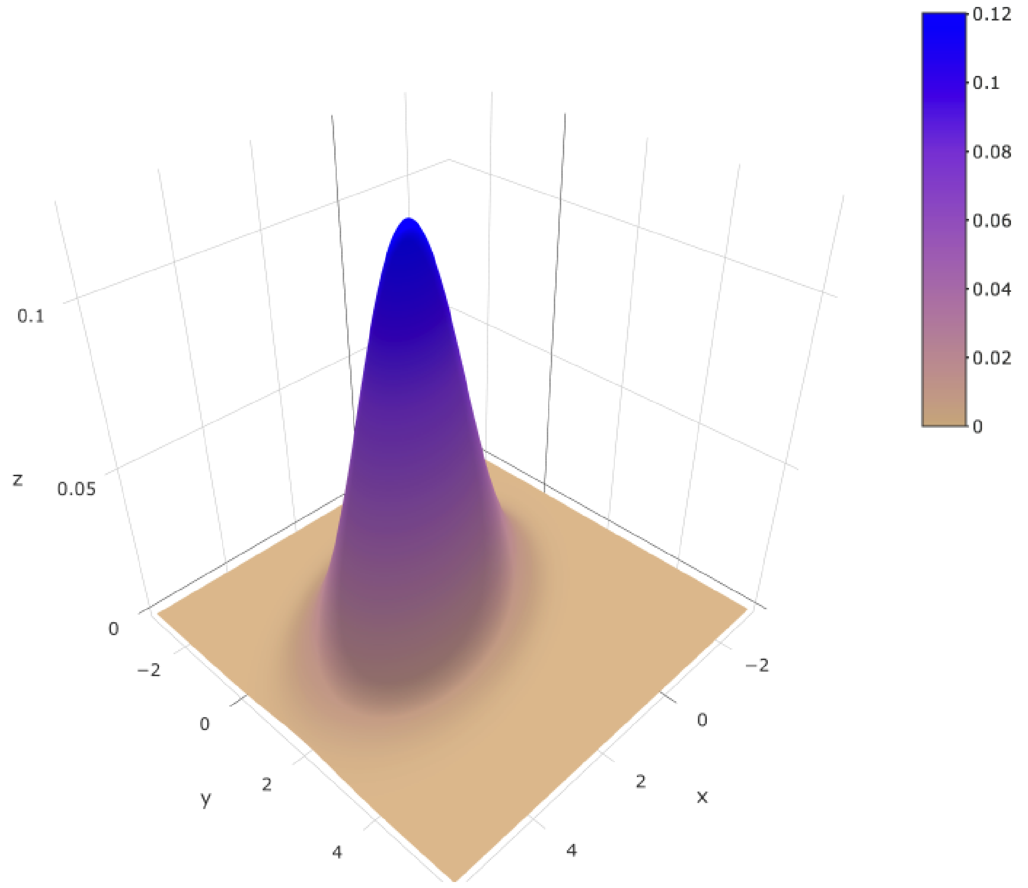
Univariate normal with mean 2 and variance 1

Density shape of a bivariate normal



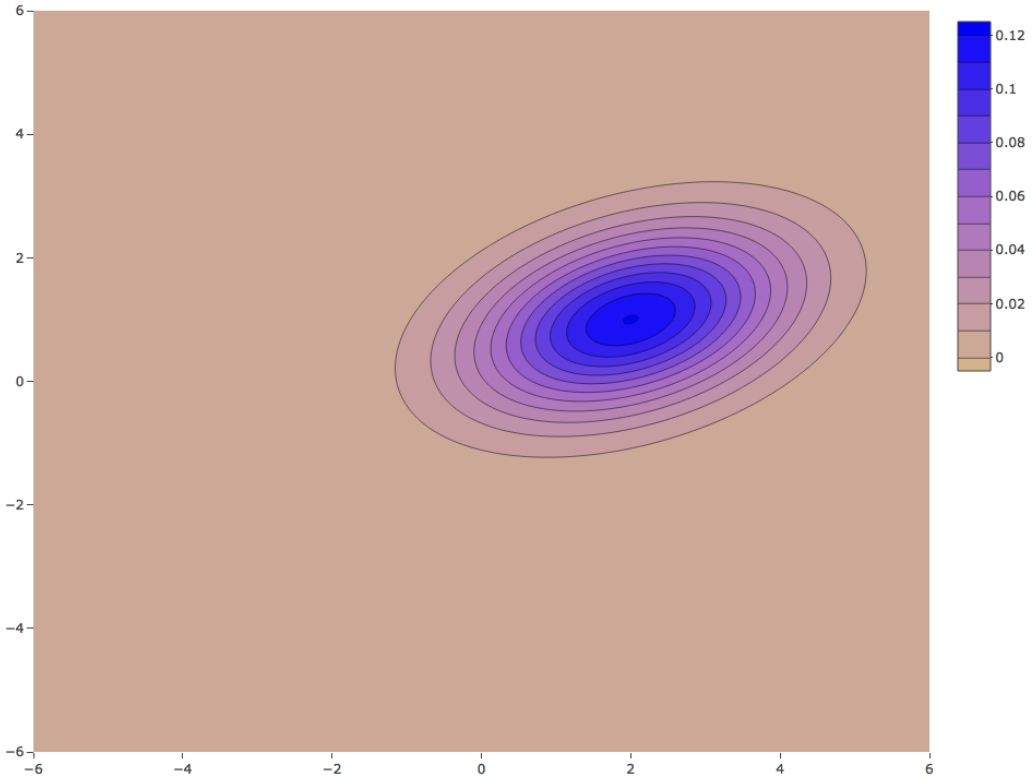
Bivariate normal density - 3D density plot

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$



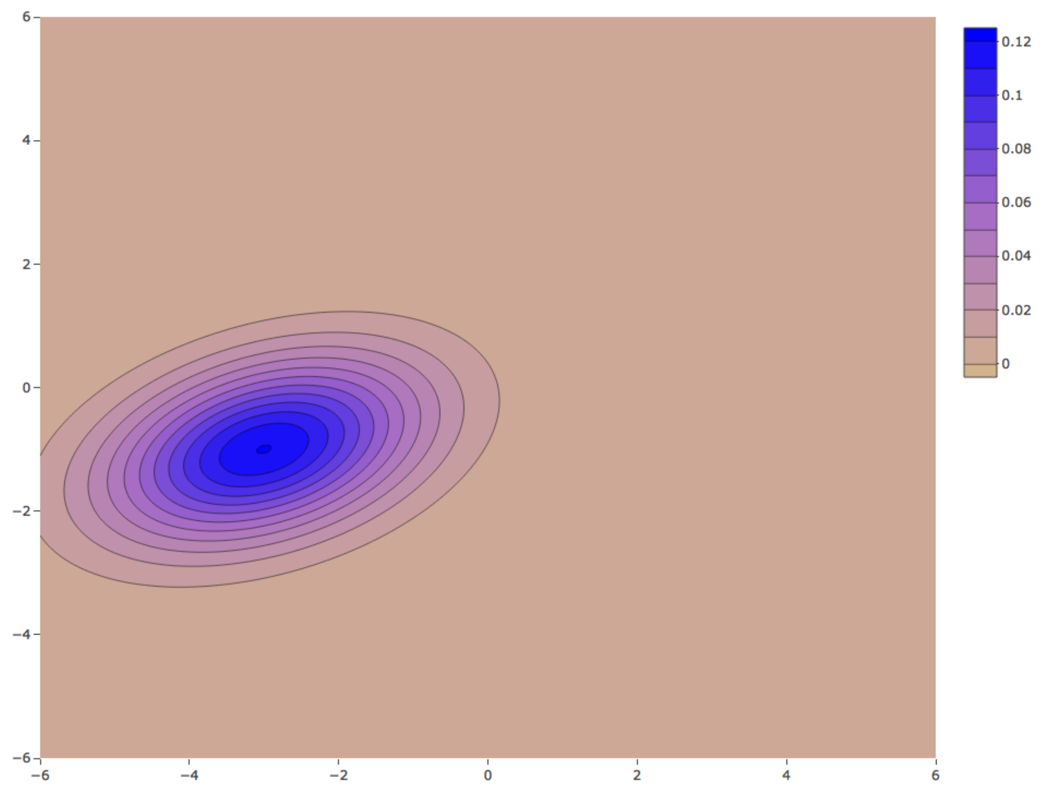
Bivariate normal density - contour plot

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$



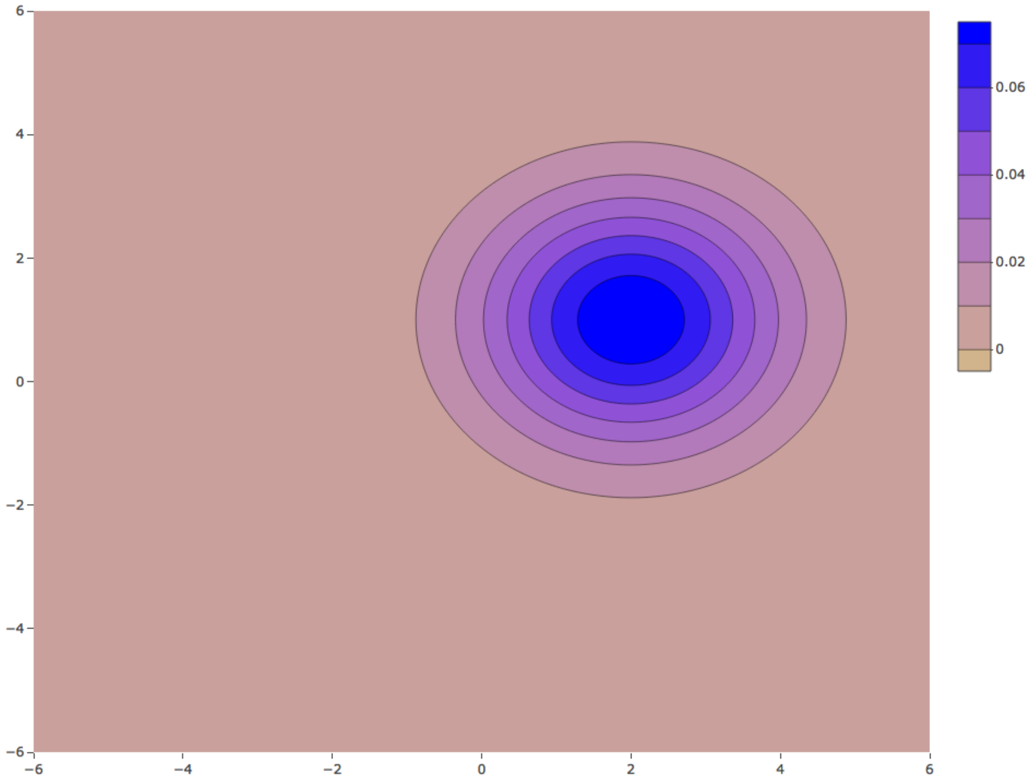
Bivariate normal density with a different mean

$$\mu = \begin{pmatrix} -1 \\ -3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$



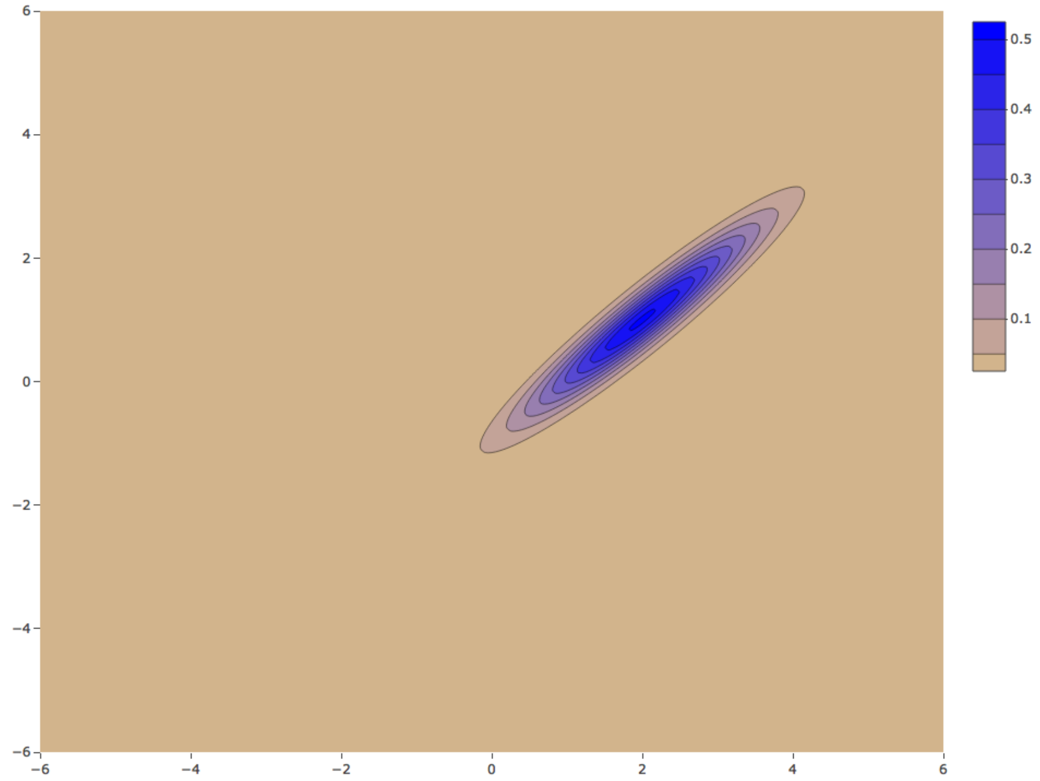
Bivariate normal density with a different variance

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



Bivariate normal density with strong correlation

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$$



Functions for statistical distributions in R

Normal	Multivariate Normal	t	Multivariate t
`rnorm`	`rmvnorm`	`rt`	`rmvt`
`dnorm`	`dmvnorm`	`dt`	`dmvt`
`pnorm`	`pmvnorm`	`pt`	`pmvt`
`qnorm`	`qmvnorm`	`qt`	`qmvt`

Functions for statistical distributions in R

Normal	Multivariate Normal	t	Multivariate t
rnorm	rmvnorm	rt	rmvt
dnorm	dmvnorm	dt	dmvt
pnorm	pmvnorm	pt	pmvt
qnorm	qmvnorm	qt	qmvt

- The first letter denotes
  - `p` for "probability"
  - `q` for "quantile"
  - `d` for "density"
  - `r` for "random"
- Followed by the distribution name
  - `norm`
  - `mvnorm`
  - `t`
  - `mvt`

The rmvnorm function

```
library(mvtnorm)
rmvnorm(n, mean , sigma)
```

Need to specify:

- `n` the number of samples
- `mean` the mean of the distribution
- `sigma` the variance-covariance matrix

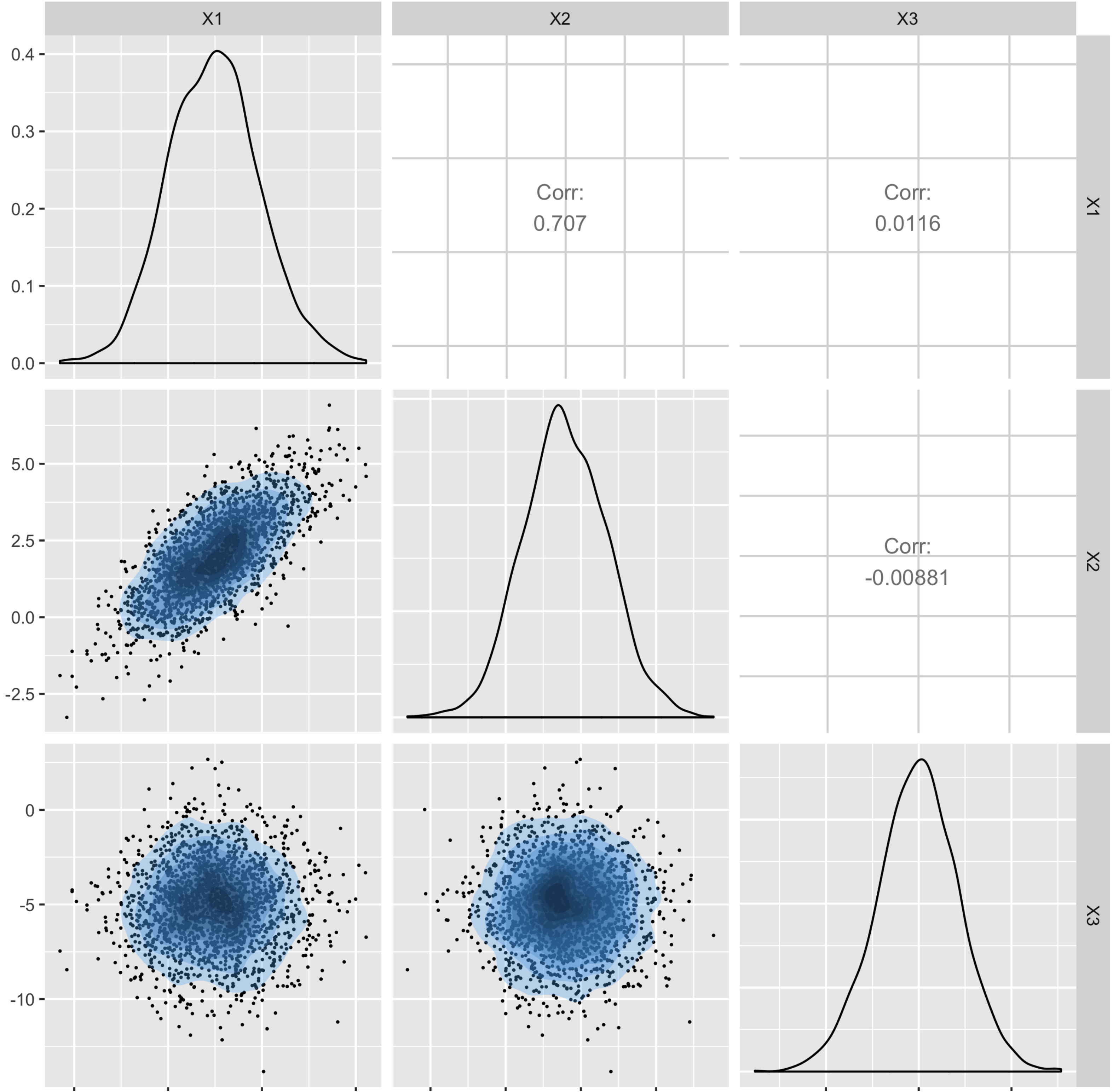
Using rmvnorm to generate random samples

Generate 1000 samples from a 3 dimensional normal with

$$\mu = \begin{pmatrix} 1 \\ 2 \\ -5 \end{pmatrix} \Sigma = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

```
mu1 <- c(1, 2, -5)
sigma1 <- matrix(c(1,1,0,
                   1,2,0,
                   0,0,5),3,3)
set.seed(34)
rmvnorm(n = 1000, mean = mu1, sigma = sigma1)
```

Plot of generated samples



MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

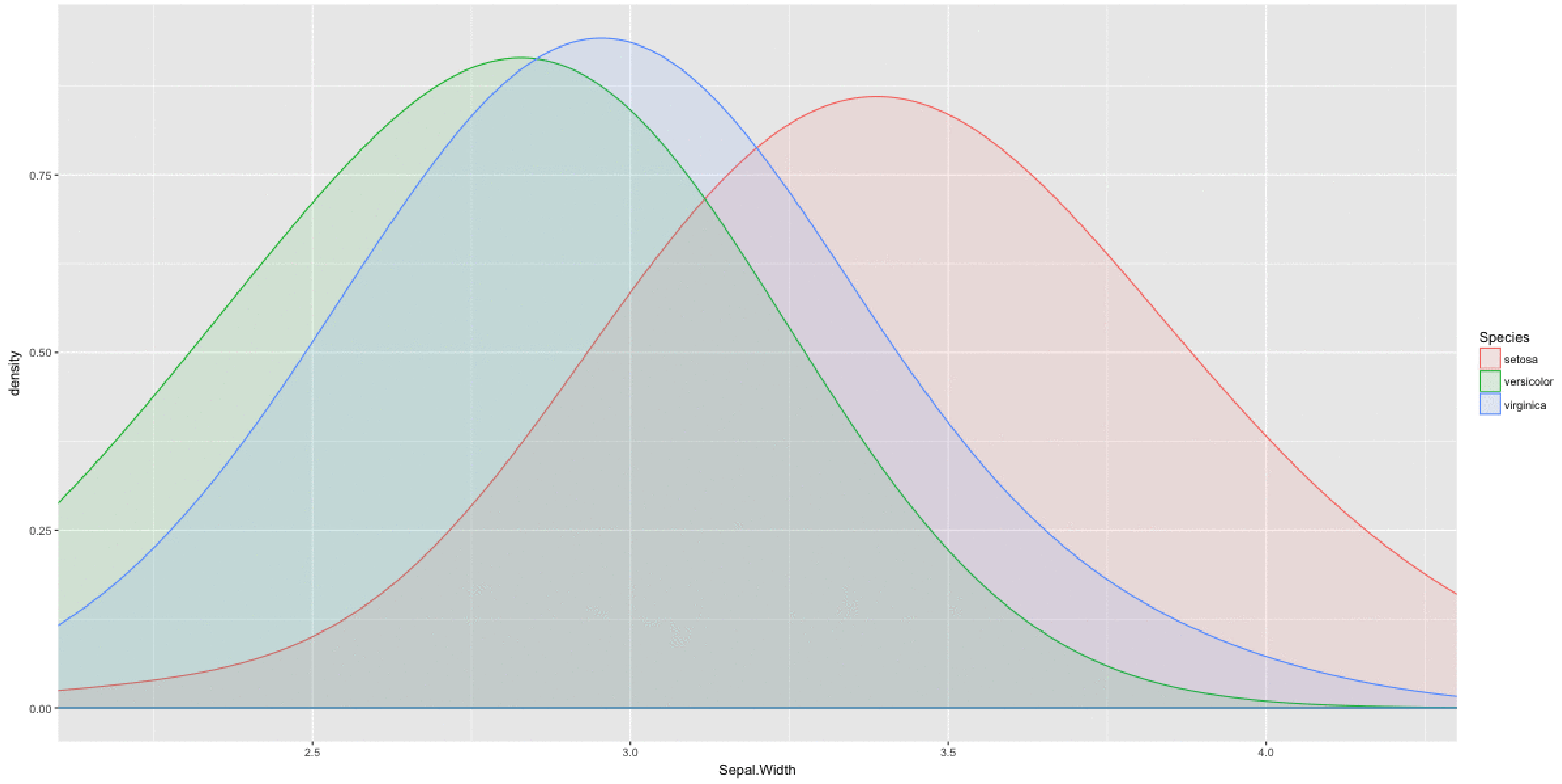
Let's practice simulating from a multivariate normal distribution!

MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

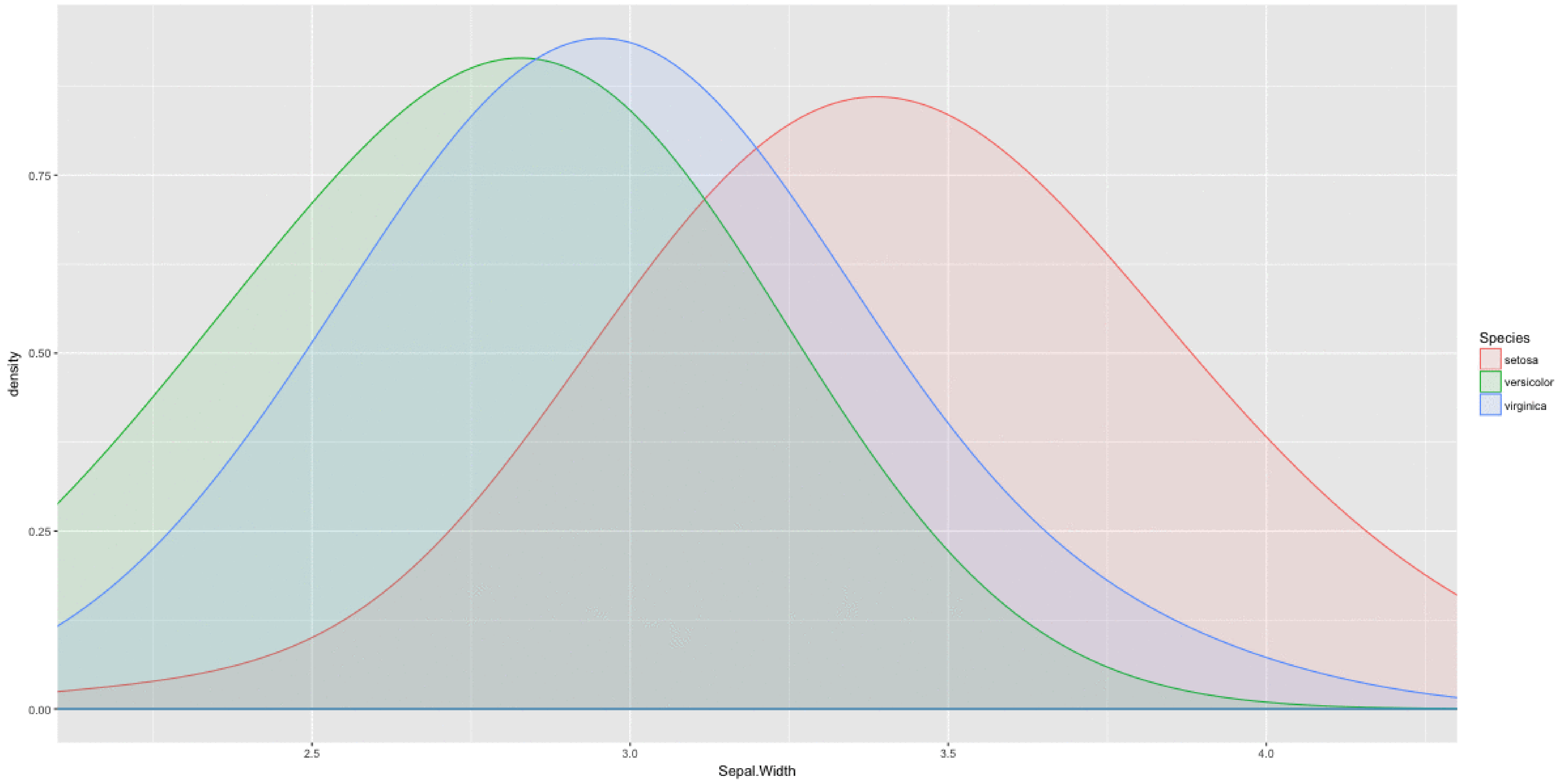
Density of a multivariate normal distribution

Surajit Ray
Senior Lecturer, University of Glasgow

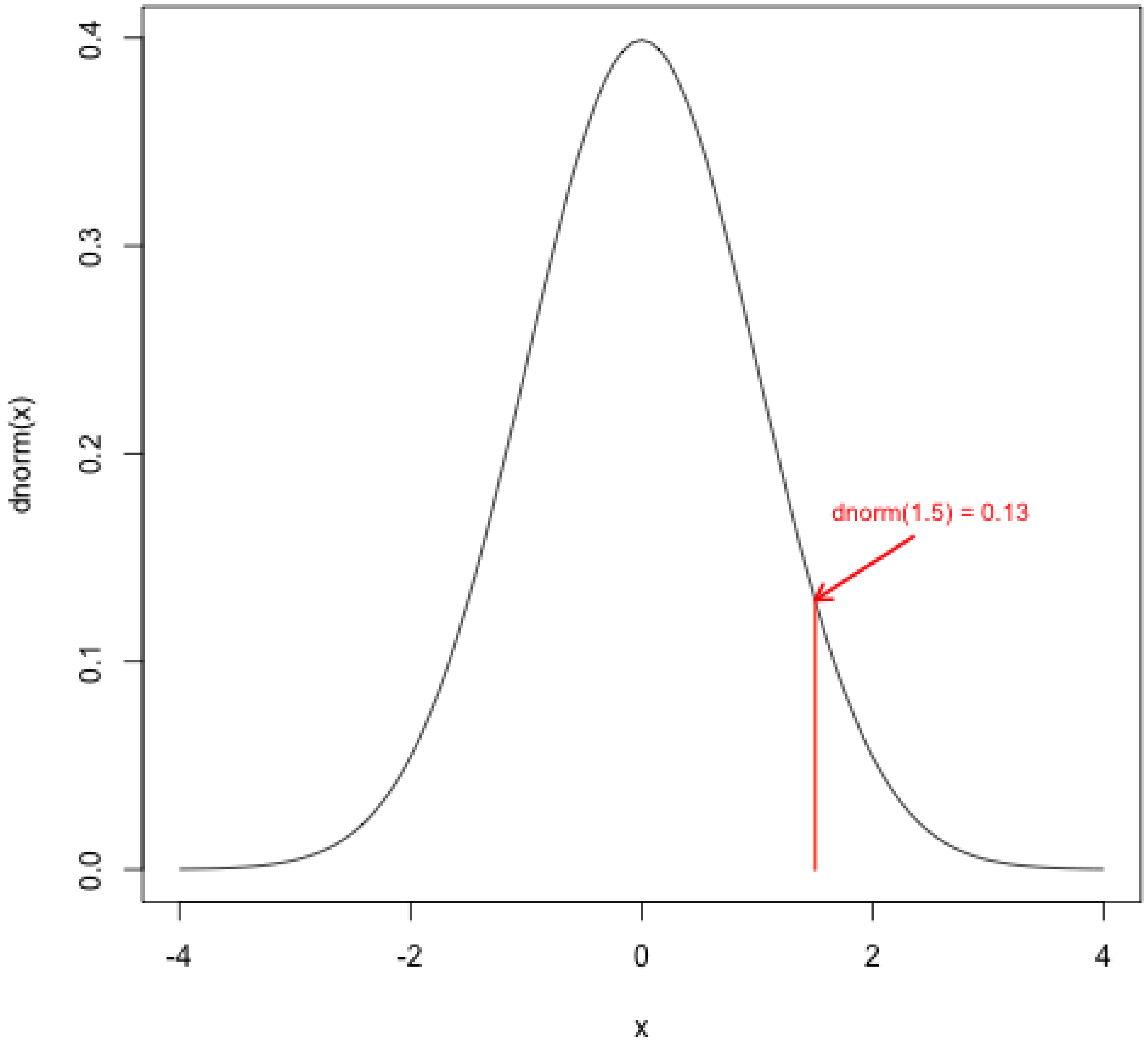
Why calculate the density of a distribution?



Why calculate the density of a distribution?



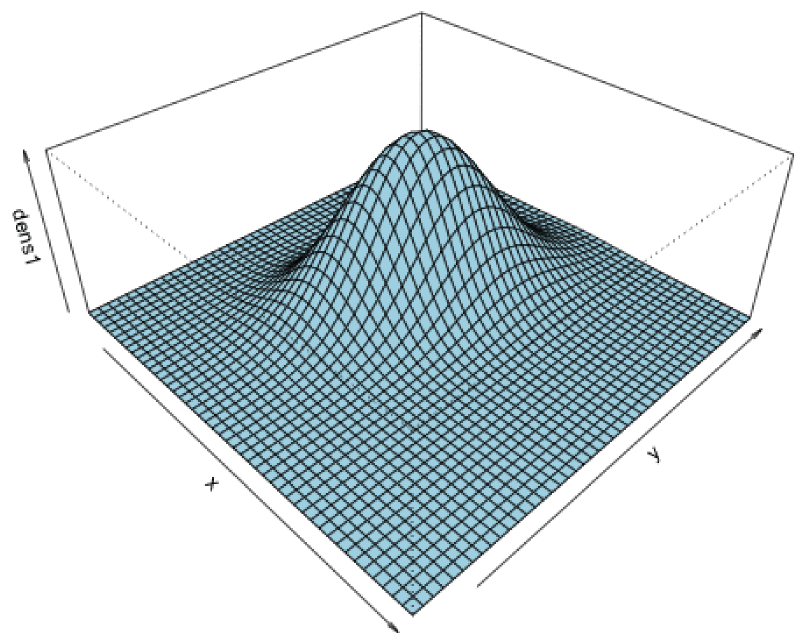
Univariate normal functions `dnorm()`



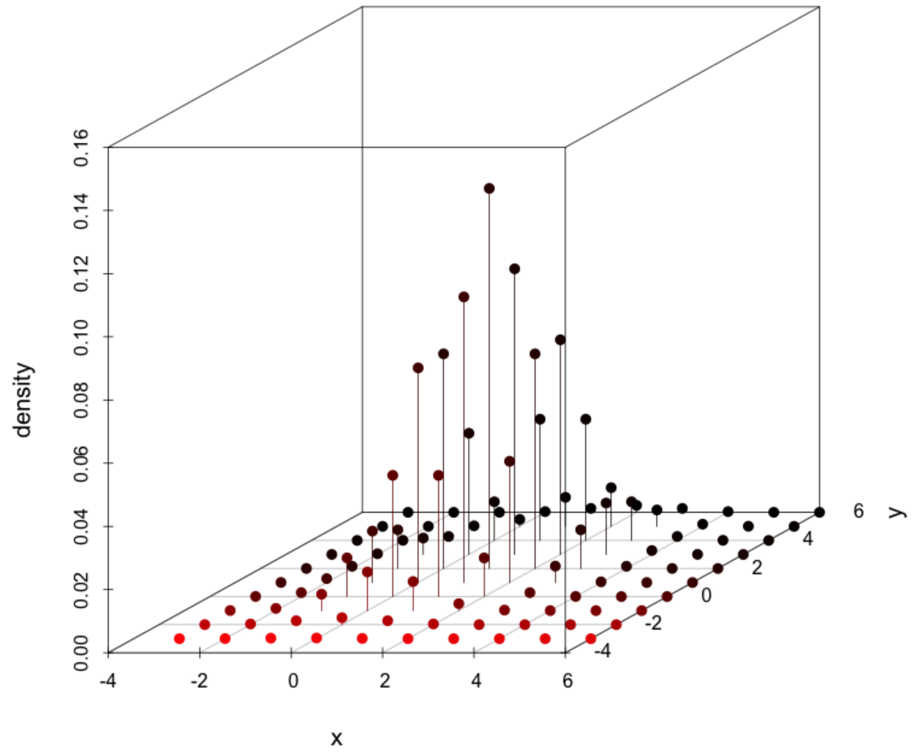
Probability density of a bivariate normal

Standard bivariate normal with

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



Density heights calculated at several locations (xy coordinates)



Density using dmvnorm

```
library(mvtnorm)
dmvnorm(x, mean, sigma)
```

- `x` can be a row vector or a matrix

```
mu1 <- c(1, 2)
sigma1 <- matrix(c(1, .5, .5, 2), 2)
dmvnorm(x = c(0, 0), mean = mu1, sigma = sigma1)
```

```
0.0384
```

Density at multiple points using dmvnorm

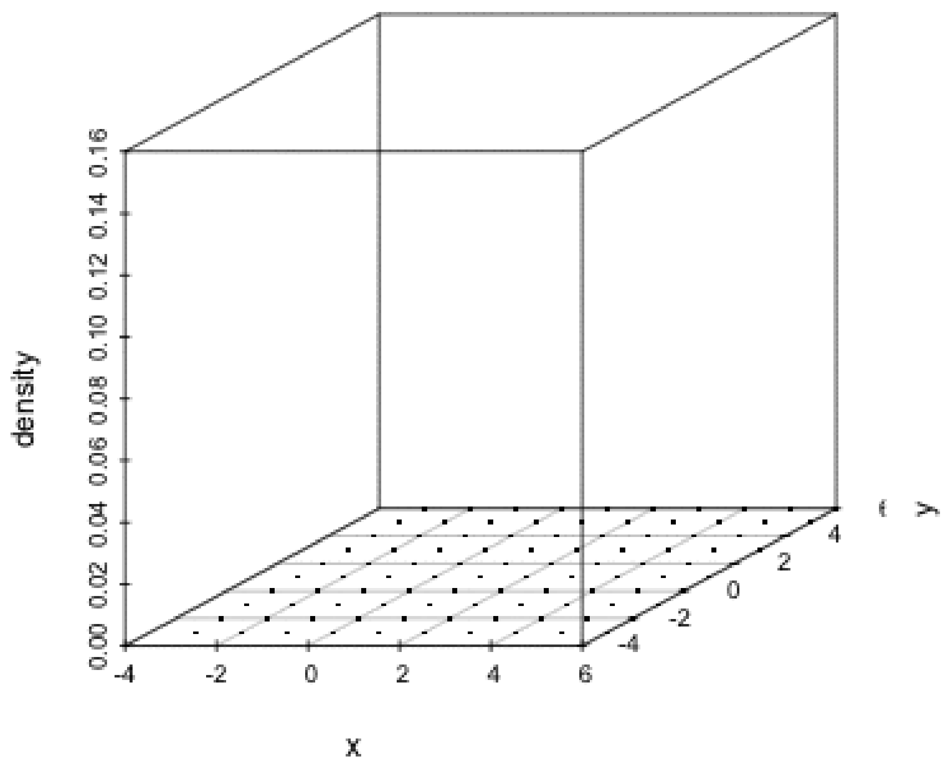
```
x <- rbind(c(0, 0), c(1, 1), c(0, 1)); x

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

```
dmvnorm(x = x, mean = mu, sigma = sigma)

[1] 0.0384 0.0904 0.0679
```

Plotting bivariate densities with perspective plot

Steps:

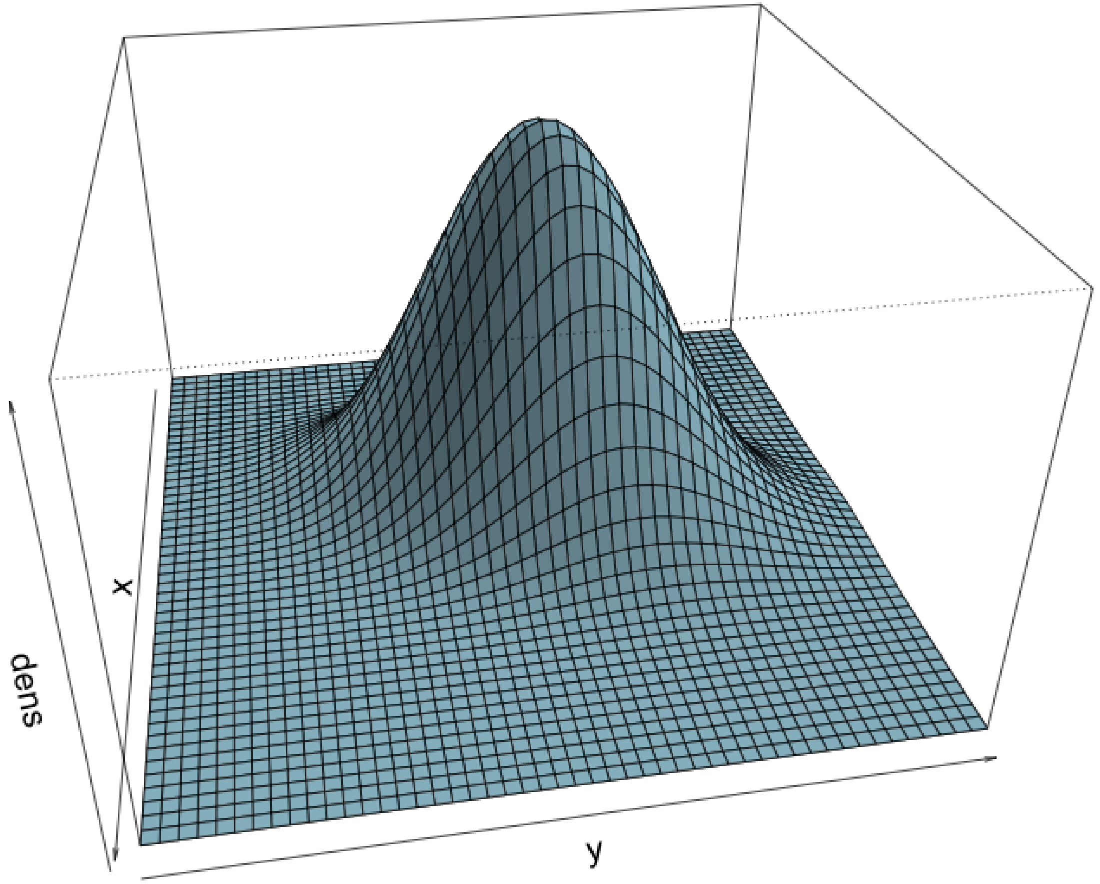
- Create grid of x and y coordinates
- Calculate density on grid



Plotting bivariate densities with perspective plot

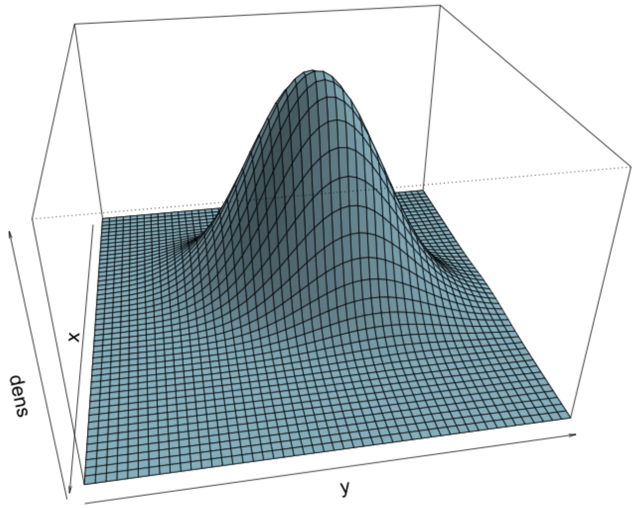
Steps:

- Create grid of x and y coordinates
- Calculate density on grid
- Convert densities into a matrix
- Create perspective plot using `persp()` function



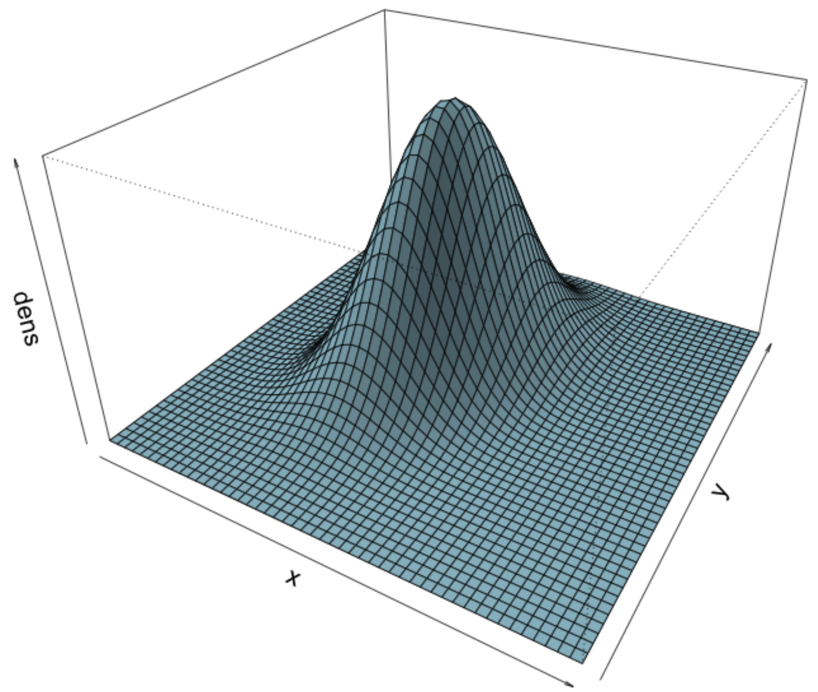
Code for plotting bivariate densities

```
# Create grid
d <- expand.grid(seq(-3, 6, length.out = 50 ), seq(-3,  6, length.out = 50))

# Calculate density on grid
dens1 <- dmvnorm(as.matrix(d), mean=c(1,2), sigma=matrix(c(1, .5, .5, 2), 2))

# Convert to matrix
dens1 <-  matrix(dens1, nrow = 50 )

# Use perspective plot
persp(dens1, theta = 80, phi = 30, expand = 0.6, shade = 0.2,
        col = "lightblue", xlab = "x", ylab = "y", zlab = "dens")
```

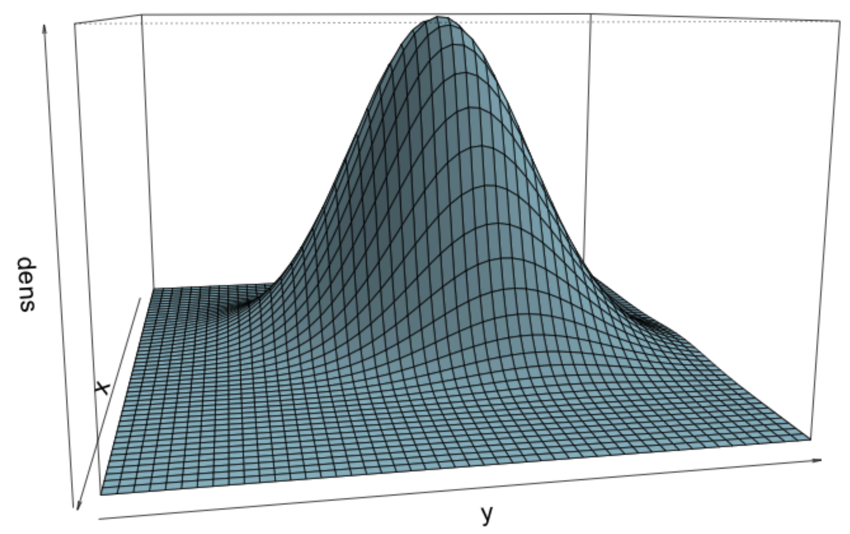


Changing viewing angle in perspective plot

`persp()` with `theta = 30, phi = 30`



`persp()` with `theta = 80, phi = 10`



MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Let's practice!

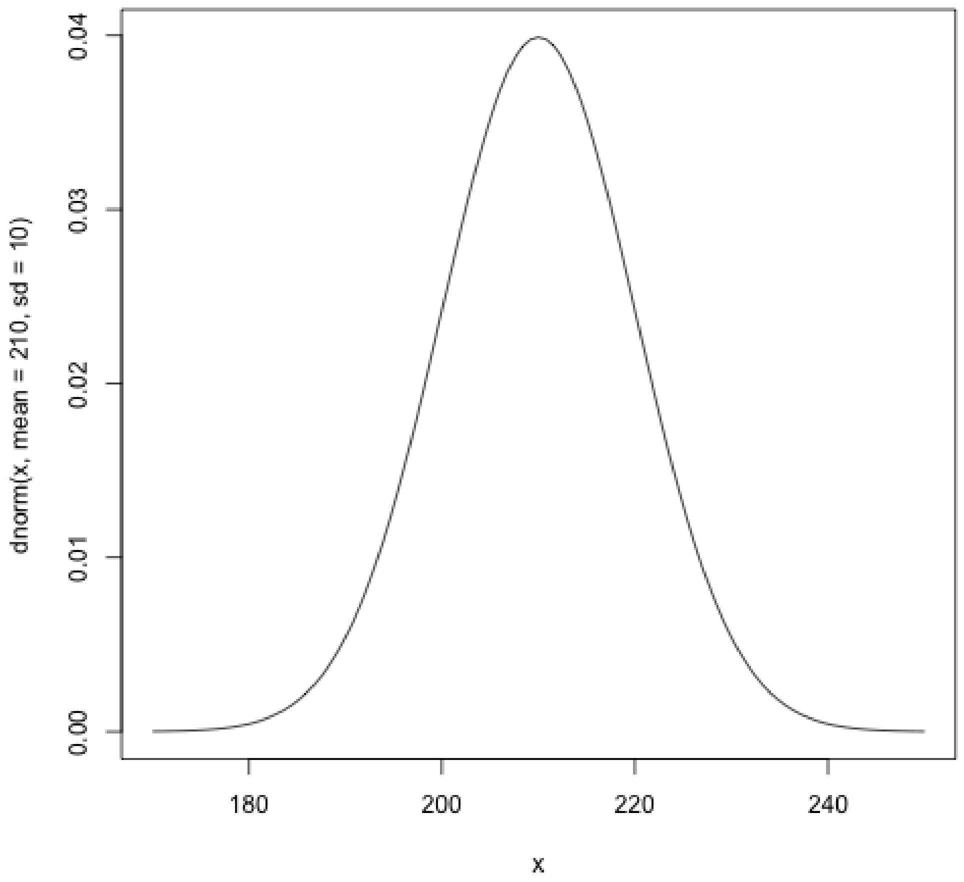
MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Cumulative Distribution and Inverse CDF

Surajit Ray
Senior Lecturer, University of Glasgow

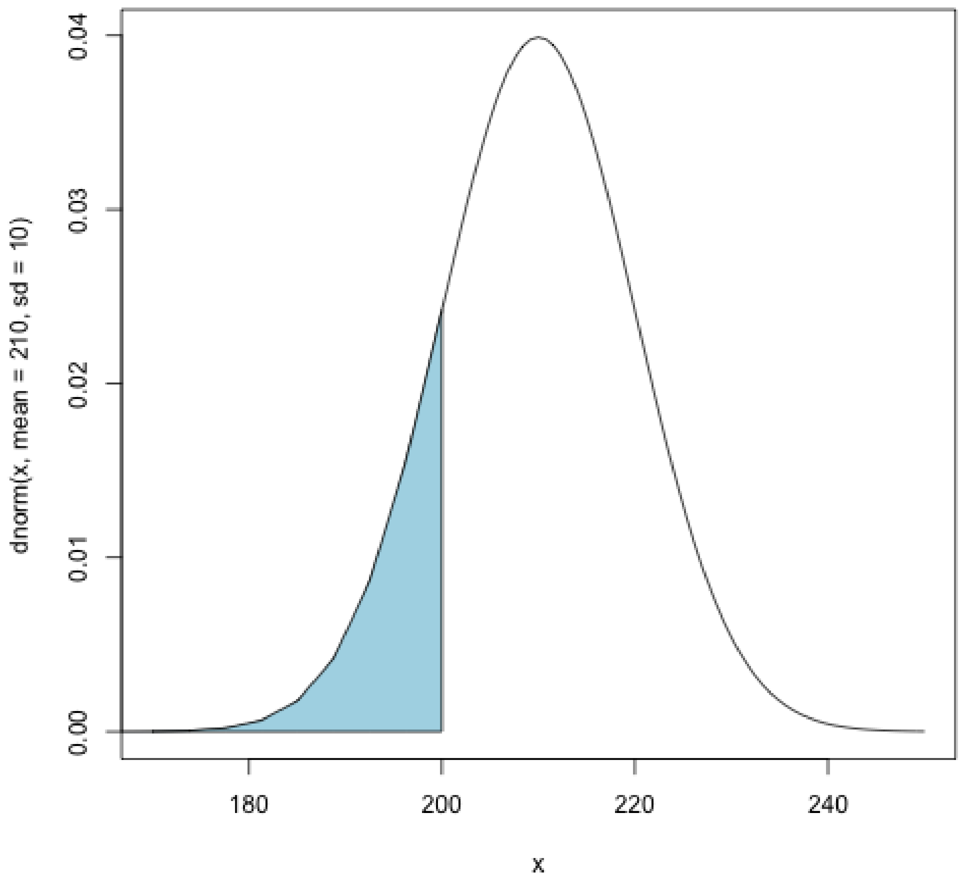
When do we need to calculate CDF and inverse CDF?

When do we need to calculate CDF and inverse CDF?



Normal density with $\mu = 210$ and $\sigma = 10$

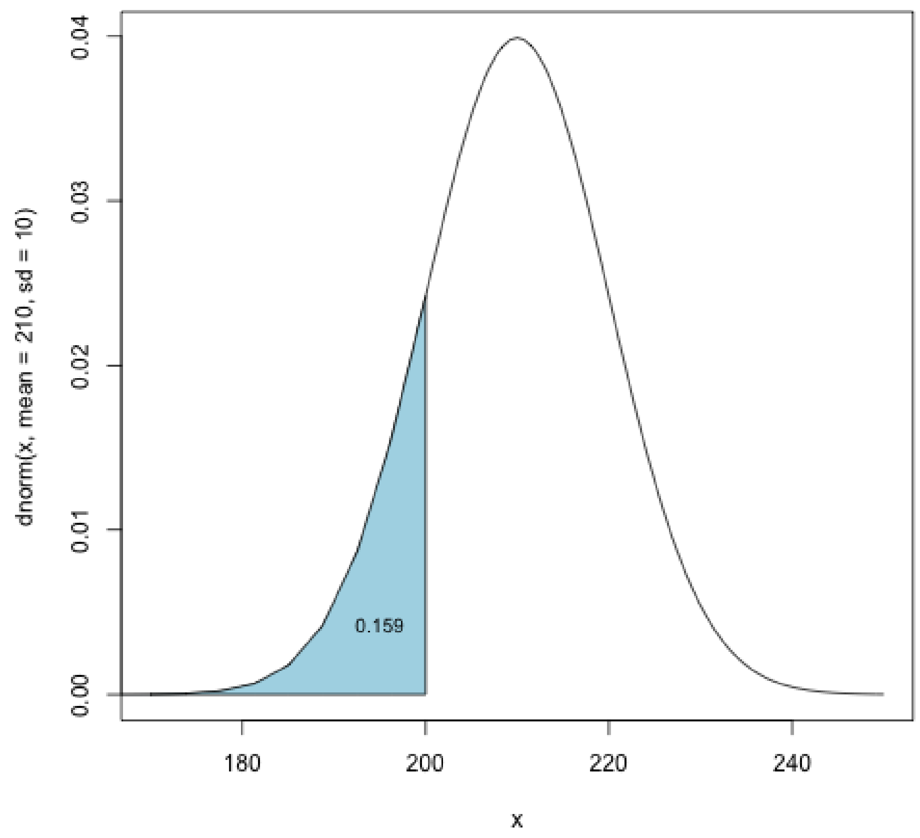
When do we need to calculate CDF and inverse CDF?



Area under the curve for $x < 200$

When do we need to calculate CDF and inverse CDF?

```
pnorm(200, mean = 210, sd = 10)
[1] 0.159
```

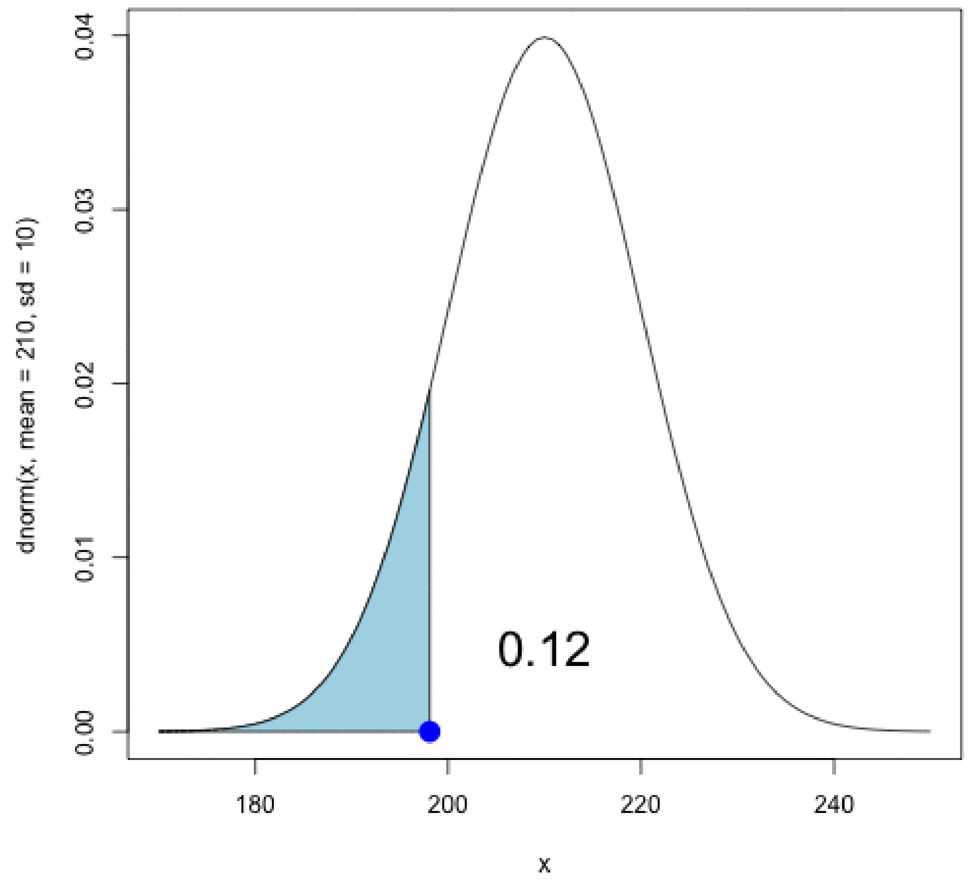


When do we need to calculate CDF and inverse CDF?

What is the x_0 such that the cumulative probability at x_0 is 0.95?

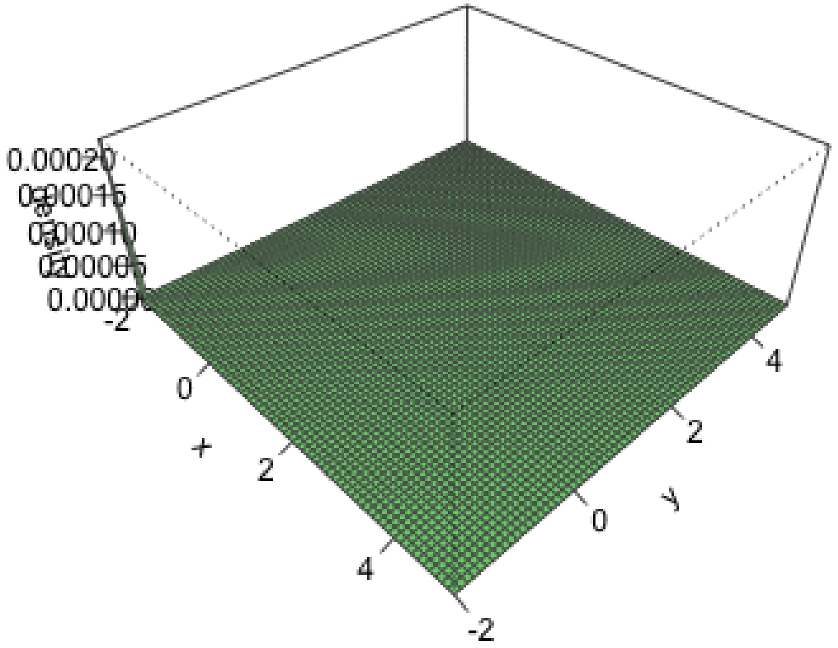
```
qnorm( p = 0.95, mean = 210, sd = 10)
[1] 226.45
```

$\Rightarrow$ 95% of the coffee jars will have less than **226.45** grams of coffee



Cumulative distribution for a bivariate normal

Bivariate CDF at x = 2 and y = 4 for a normal with $\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 2 \end{pmatrix}$



Cumulative distribution using pmvnorm

Bivariate CDF at x = 2 and y = 4 for a normal with $\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$

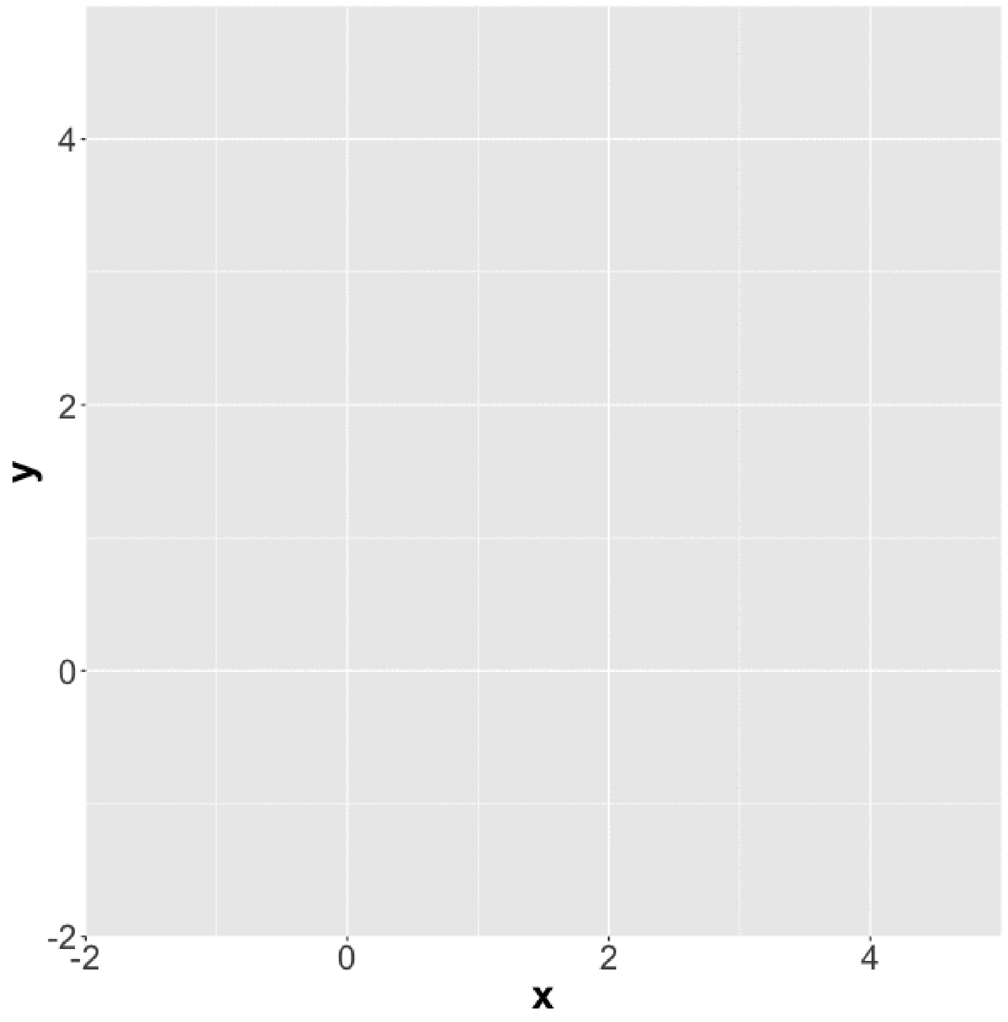
```
mu1 <- c(1, 2)
sigma1 <- matrix(c(1, 0.5, 0.5, 2), 2)

pmvnorm(upper = c(2, 4), mean = mu1, sigma = sigma1)
```

```
[1] 0.79
attr(,"error")
[1] 1e-15
attr(,"msg")
[1] "Normal Completion"
```

Probability between two values using pmvnorm

Probability of $1 < x < 2$ and $2 < y < 4$

```
pmvnorm(lower = c(1, 2),
          upper = c(2, 4),
             mean = mu1,
              sigma = sigma1)
```

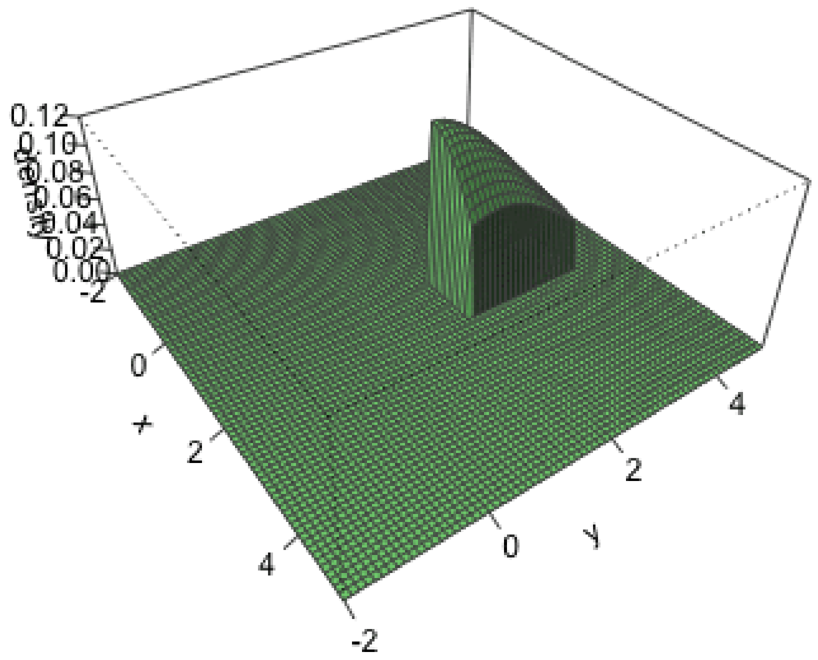


Probability between two values using pmvnorm

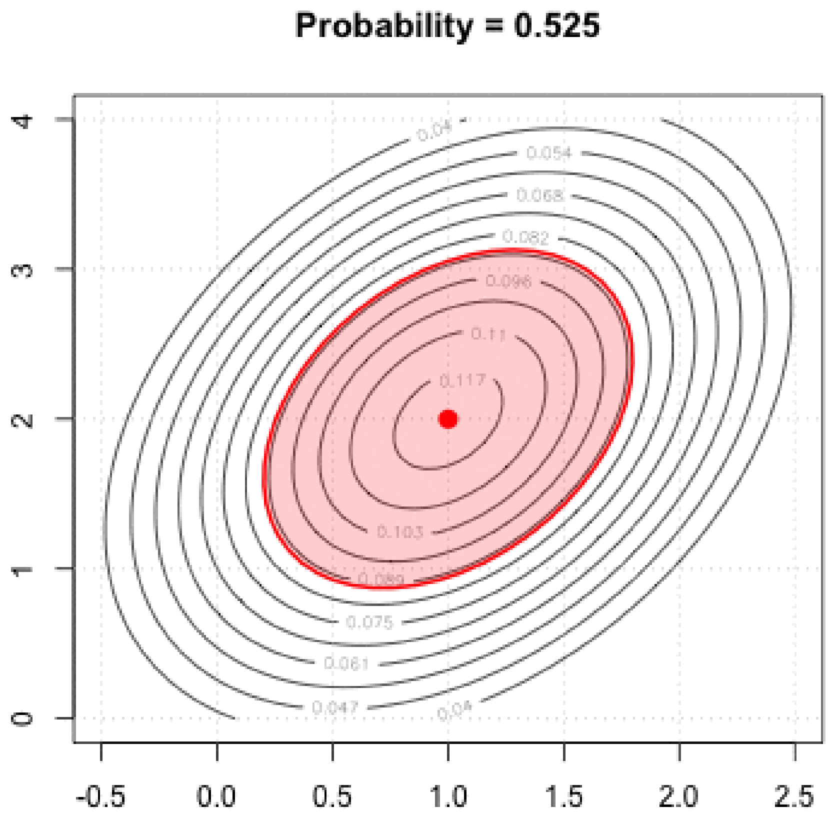
Probability of $1 < x < 2$ and $2 < y < 4$

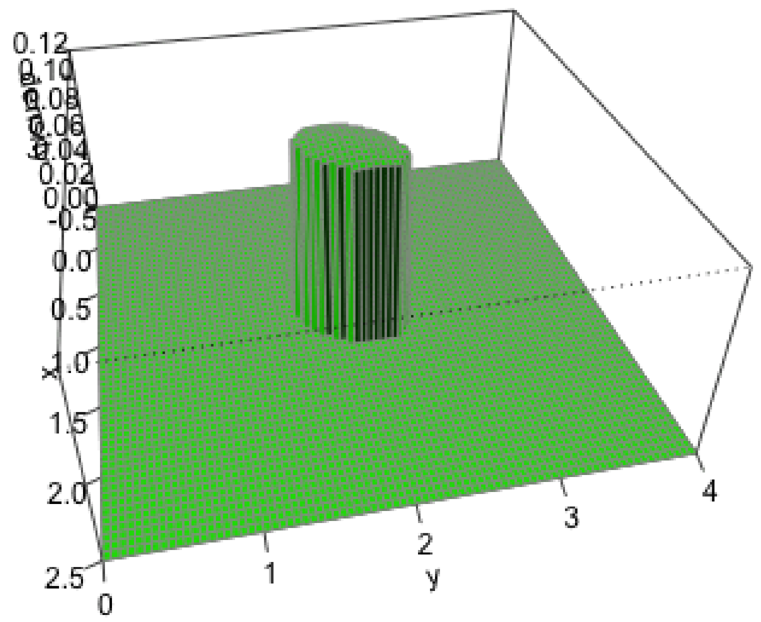
```
pmvnorm(lower = c(1, 2),
          upper = c(2, 4),
             mean = mu1,
              sigma = sigma1)
```

```
[1] 0.163
```



Inverse CDF for bivariate normal





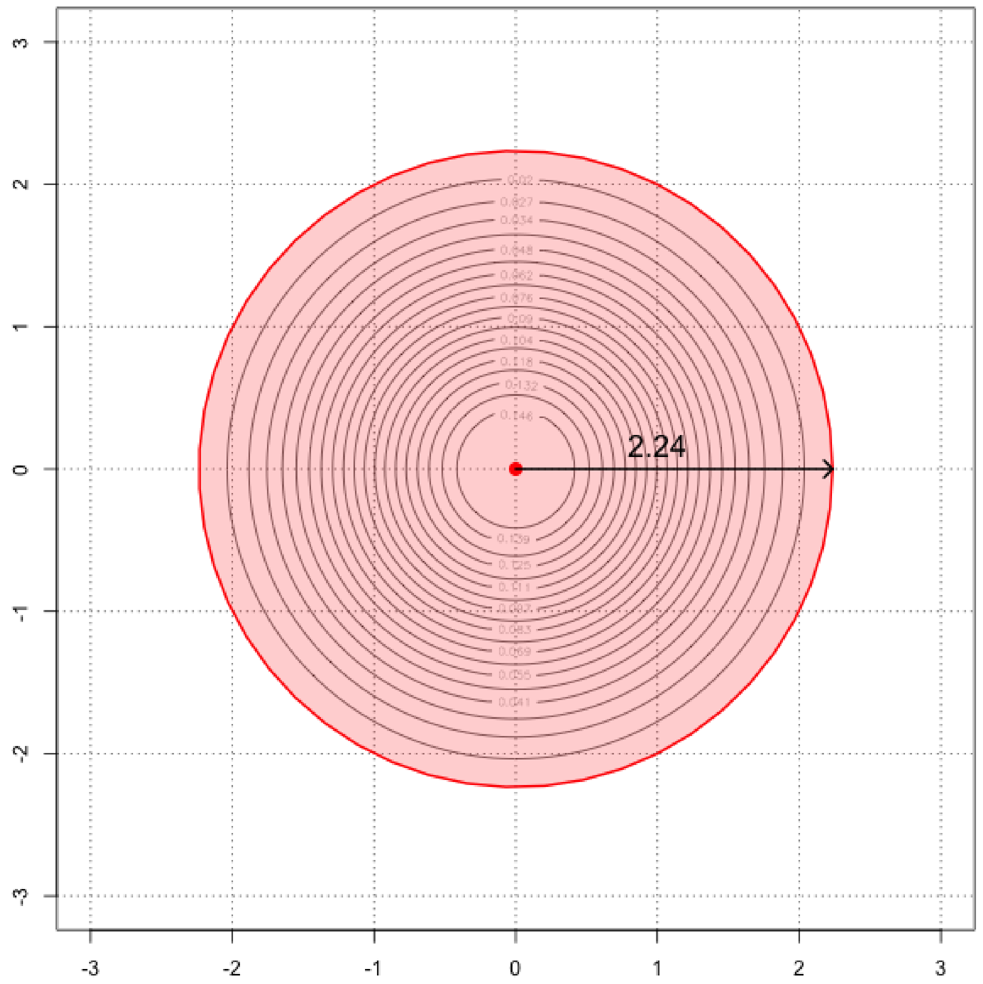
Dark red ellipse is the 0.95 quantile

Implementing qmvnorm to calculate quantiles

```
sigma1 <- diag(2)
sigma1
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
qmvnorm(p = 0.95, sigma = sigma1,
                      tail = "both")
```

```
$quantile
[1] 2.24

$f.quantile
[1] -1.31e-06

attr(,"message")
[1] "Normal Completion"
```



The red circle with radius 2.24 contains 0.95 of the probability

MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Let's practice!

MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Checking normality of multivariate data

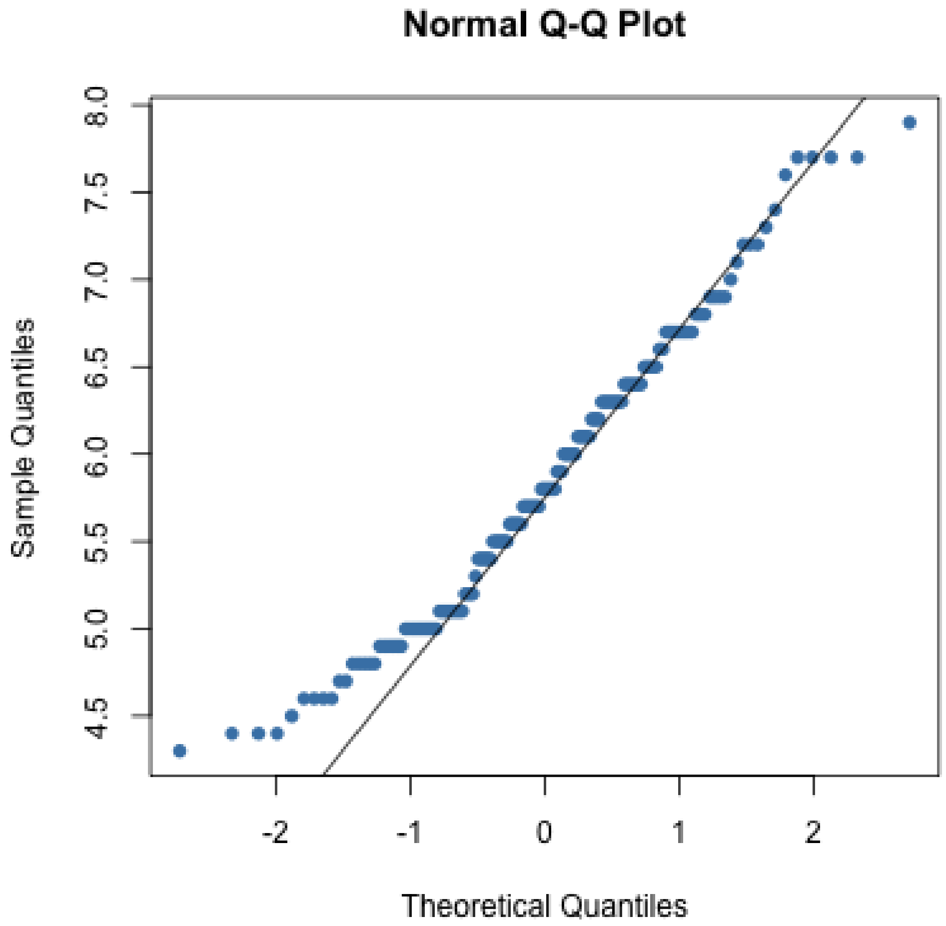
Surajit Ray
Senior Lecturer, University of Glasgow

Why check normality?

- Classical statistical techniques that assume univariate/multivariate normality:
  - Multivariate regression
  - Discriminant analysis
  - Model-based clustering
  - Principal component analysis (PCA)
  - Multivariate analysis of variance (MANOVA)

Review: univariate normality tests

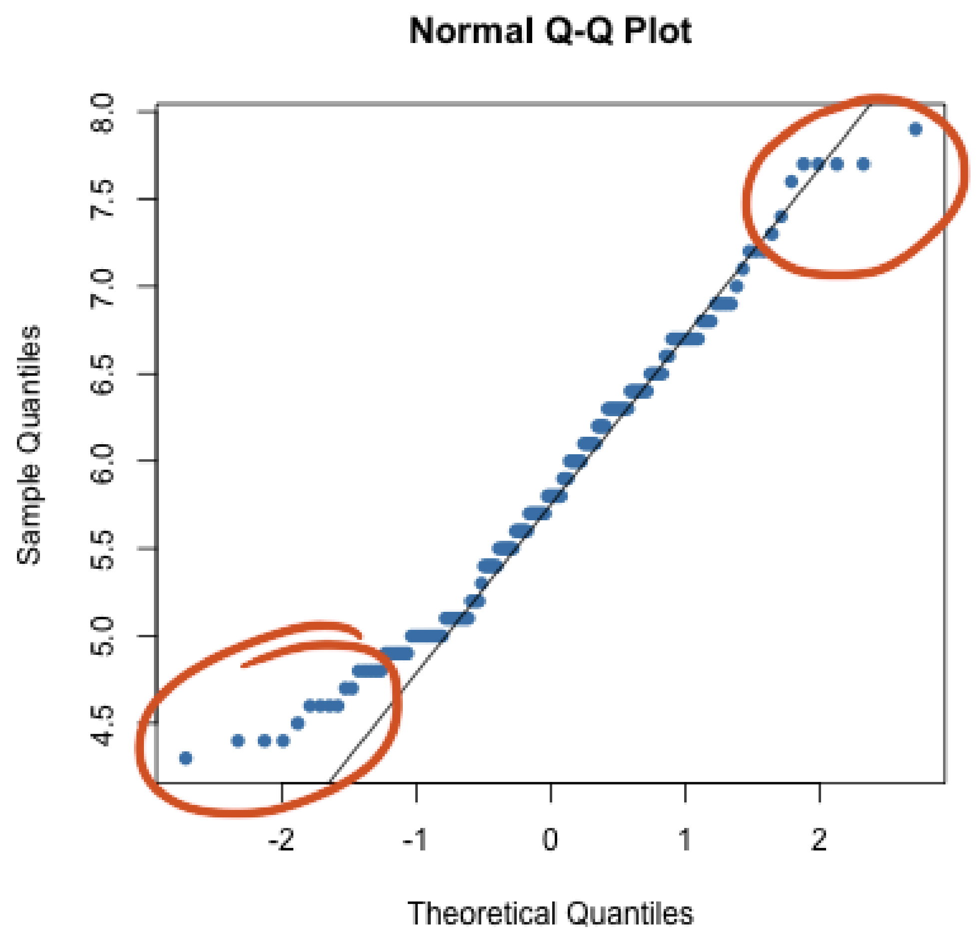
```
qqnorm(iris_raw[, 1])
qqline(iris_raw[, 1])
```



- If the values lie along the reference line the distribution is close to normal

Review: univariate normality tests

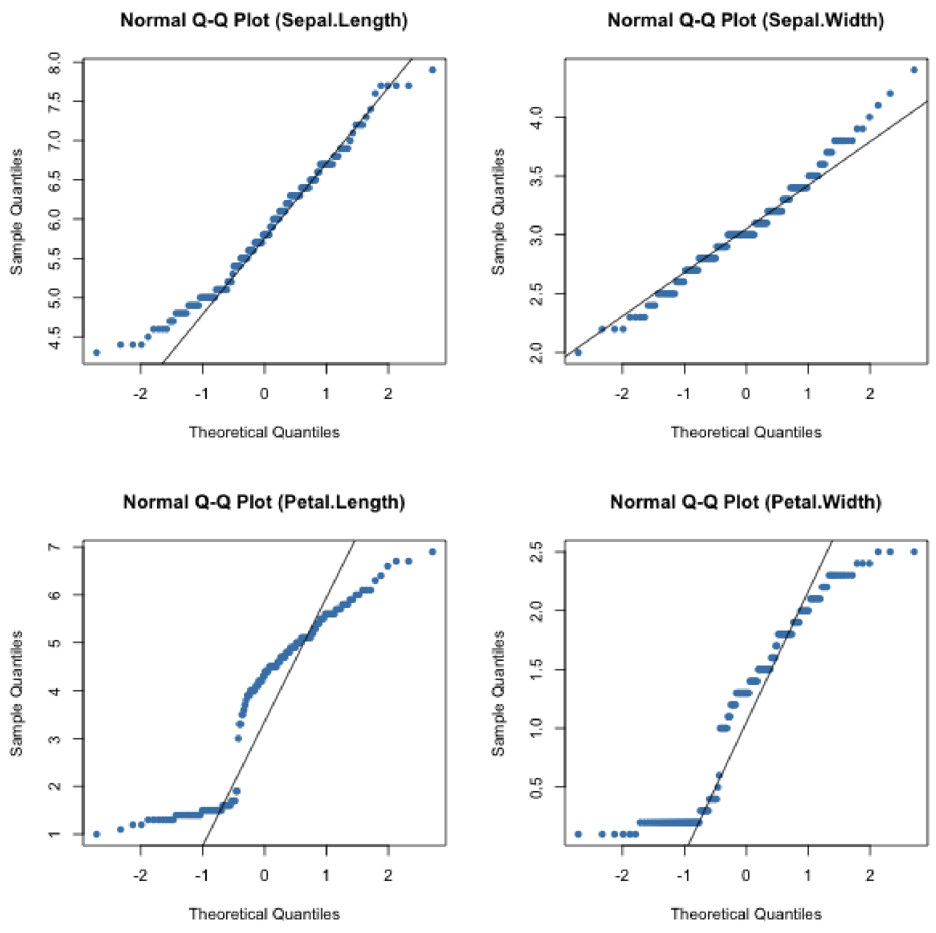
```
qqnorm(iris_raw[, 1])
qqline(iris_raw[, 1])
```



- If the values lie along the reference line the distribution is close to normal
- Deviation from the line might indicate
  - heavier tails
  - skewness
  - outliers
  - clustered data

qqnorm of all variables

```
uniPlot(iris_raw[, 1:4])
```



MVN library multivariate normality test functions

- Multivariate normality tests by
  - Mardia
  - Henze-Zirkler
  - Royston
- Graphical appoaches
  - chi-square Q-Q
  - perspective
  - contour plots

MVN library multivariate normality test functions

- Multivariate normality tests by
  - Mardia ✓
  - Henze-Zirkler ✓
  - ~~Royston~~
- Graphical appoaches
  - chi-square Q-Q ✓
  - ~~perspective~~
  - ~~contour plots~~

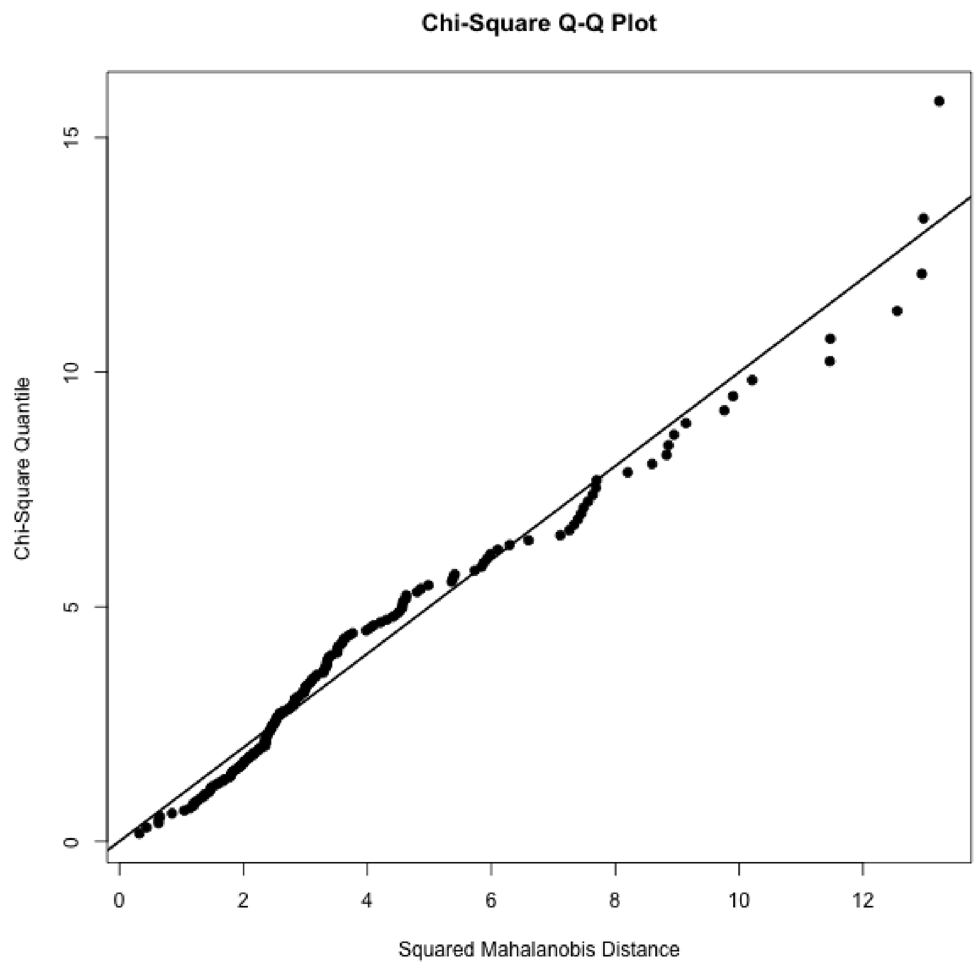
Using mardiaTest to check multivariate normality

```
mardiaTest(iris_raw[, 1:4])
```

```
   Mardia Multivariate Normality Test
---------------------------------------
   data : iris_raw[, 1:4]

    g1p            : 2.697
   chi.skew        : 67.43
   p.value.skew    : 4.758e-07

   g2p             : 23.74
   z.kurtosis      : -0.2301
   p.value.kurt    : 0.818

   chi.small.skew : 69.33
   p.value.small  : 2.342e-07

   Result          : Data are not multivariate normal.
-------------------------------------------------------
```

Using qqplot from mardiaTest to check multivariate normality

```
mardiaTest(iris_raw[, 1:4], qqplot = TRUE)
```



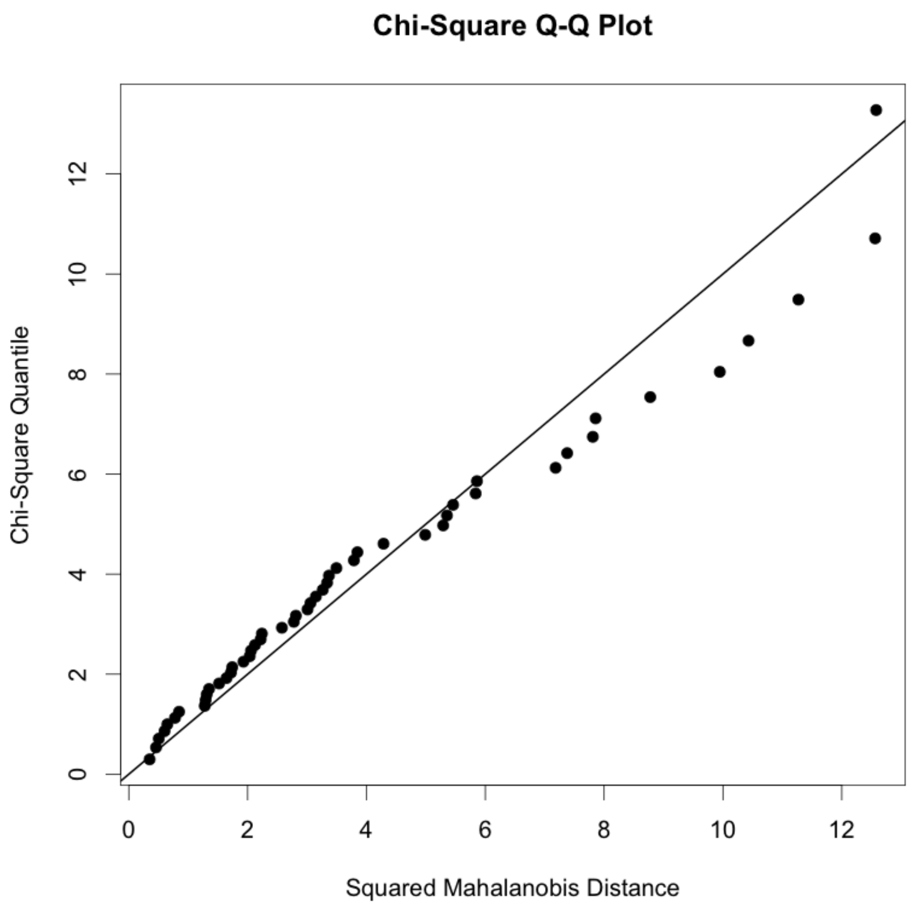
Using hzTest to check multivariate normality

```
hzTest(iris_raw[,1:4])
```

```


  Henze-Zirkler's Multivariate Normality Test
---------------------------------------------
  data : iris_raw[, 1:4]

  HZ        : 2.333269
  p-value : 0

  Result   : Data are not multivariate normal.
---------------------------------------------
```

Testing multivariate normality by species

```
mardiaTest(iris[iris_raw$Species
                 == "setosa", 1:4])

Mardia's Multivariate Normality Test
---------------------------------------
   g1p             : 3.08
   chi.skew        : 25.7
   p.value.skew    : 0.177

   g2p             : 26.5
   z.kurtosis      : 1.29
   p.value.kurt    : 0.195

   chi.small.skew : 27.85973
   p.value.small  : 0.1127617

   Result          : Data are
                  multivariate normal.
---------------------------------------
```



MULTIVARIATE PROBABILITY DISTRIBUTIONS IN R

Let's make use of the tests for multivariate normality!