Basic Statistics for Scientists in R and python

mark doerr
institute for biochemistry
university greifswald, germany

September 4, 2016

0.1 Using R Studio

0.1.1 Windows in R Studio

- Text Editor window
- Console
- Environment window
- Help and Plots Window

0.1.2 Getting help

Help/documentation viewer
Hitting F1 on a function shows help.

0.2 Basic Data Types in R

0.2.1 vector

```
x = 1

x

## [1] 1

x[1]

## [1] 1

y = (1:10)

y

##  [1]  1  2  3  4  5  6  7  8  9 10

y[2]

## [1] 2
```

0.2.2 matrix

```
a_mtr = matrix(y, nrow=2)

a_mtr

##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    3    5    7    9
## [2,]    2    4    6    8   10
```

0.2.3 list

```
a_lst = list("A", 1)

a_lst

## [[1]]
## [1] "A"
##
## [[2]]
## [1] 1
```

0.2.4 data frame

```
x = (1:10)
my_first_data_frame_df = data.frame("x"=x, "y"=x*0.1 )

my_first_data_frame_df

##     x   y
## 1   1 0.1
## 2   2 0.2
## 3   3 0.3
## 4   4 0.4
## 5   5 0.5
## 6   6 0.6
## 7   7 0.7
## 8   8 0.8
## 9   9 0.9
## 10 10 1.0
```

This shows how to access the data of the data frame

```
my_first_data_frame_df[,1] # first column

##  [1]  1  2  3  4  5  6  7  8  9 10

my_first_data_frame_df$x  # first column by name

##  [1]  1  2  3  4  5  6  7  8  9 10

my_first_data_frame_df[1,] # first line

##   x   y
## 1 1 0.1

my_first_data_frame_df[2,2] # second element of second line

## [1] 0.2
```

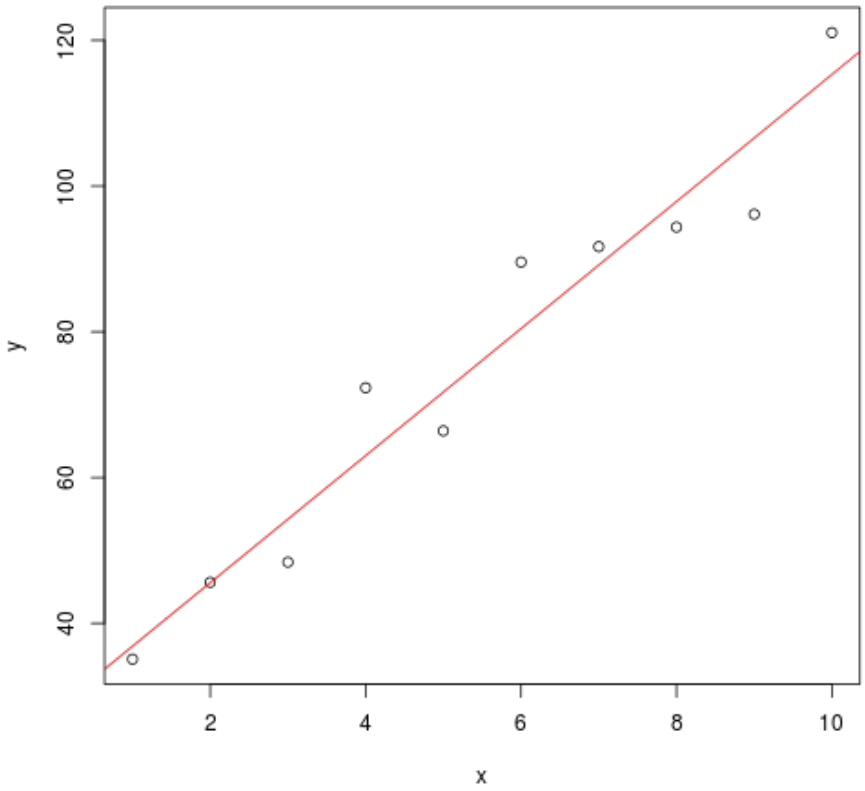
```r
x <- 1:10
w <- 20 + 10*x
w
```

```
##  [1]  30  40  50  60  70  80  90 100 110 120
```

```r
linear_sample_df <- data.frame(x=x, y=w + rnorm(10)*10)

plot(linear_sample_df)

linear_model_lm <- lm(y ~ x, data=linear_sample_df)
summary(linear_model_lm)
```

```
##
## Call:
## lm(formula = y ~ x, data = linear_sample_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4428  -4.8458  -0.8379   4.9563   9.3348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.1239     4.8138   5.842 0.000386 ***
## x             8.7173     0.7758  11.236 3.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.047 on 8 degrees of freedom
## Multiple R-squared:  0.9404,Adjusted R-squared:  0.933
## F-statistic: 126.3 on 1 and 8 DF,  p-value: 3.533e-06
```

```r
abline(linear_model_lm, col="red")
```



0.2.5 Examples of in-build data sets for testing

```
library(help = "datasets")
```

Iris

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
head(iris3)
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4
```

women

```
##   height weight
## 1     58    115
## 2     59    117
## 3     60    120
## 4     61    123
## 5     62    126
## 6     63    129
```

ELISA - DNAse

```
##   Run       conc density
## 1   1 0.04882812   0.017
## 2   1 0.04882812   0.018
## 3   1 0.19531250   0.121
## 4   1 0.19531250   0.124
## 5   1 0.39062500   0.206
## 6   1 0.39062500   0.215
```

Mean, Average, Summary

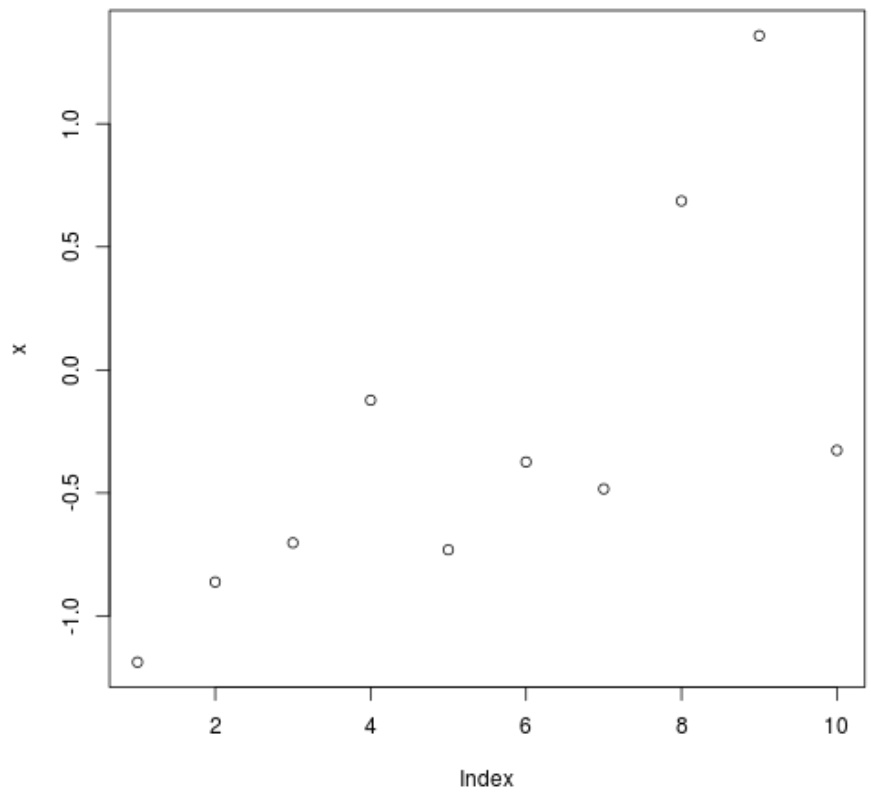
```
## [1] 60.1
## [1] 4.998889
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    51.0    57.5    61.0    60.1    62.5    69.0
## [1] 61 59 55
##
##  One Sample t-test
##
## data:  wtcsf[1:4]
## t = 31.6563, df = 3, p-value = 6.927e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  53.74326 65.75674
## sample estimates:
## mean of x
##     59.75
##
##  One Sample t-test
##
## data:  wtcsf
## t = 38.019, df = 9, p-value = 2.991e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  56.52401 63.67599
## sample estimates:
## mean of x
##      60.1
## [1] 45.5
## [1] 7.382412
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.00   40.25   46.50   45.50   49.50   59.00
```

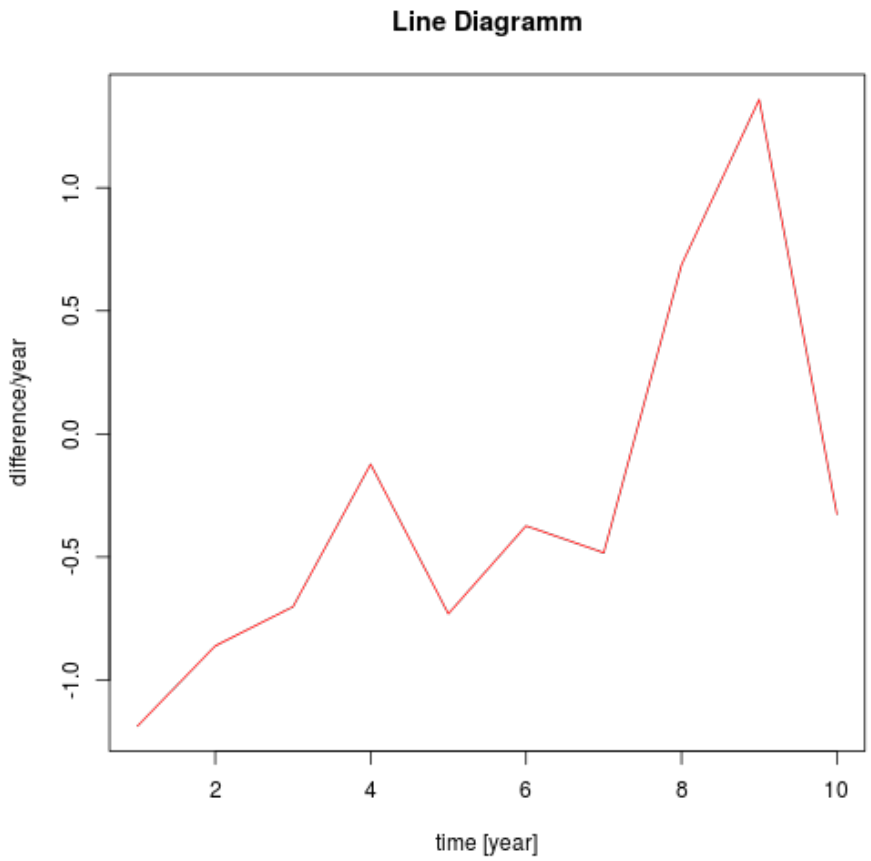
Reading/Writing Data from a File

Basic Plotting in R

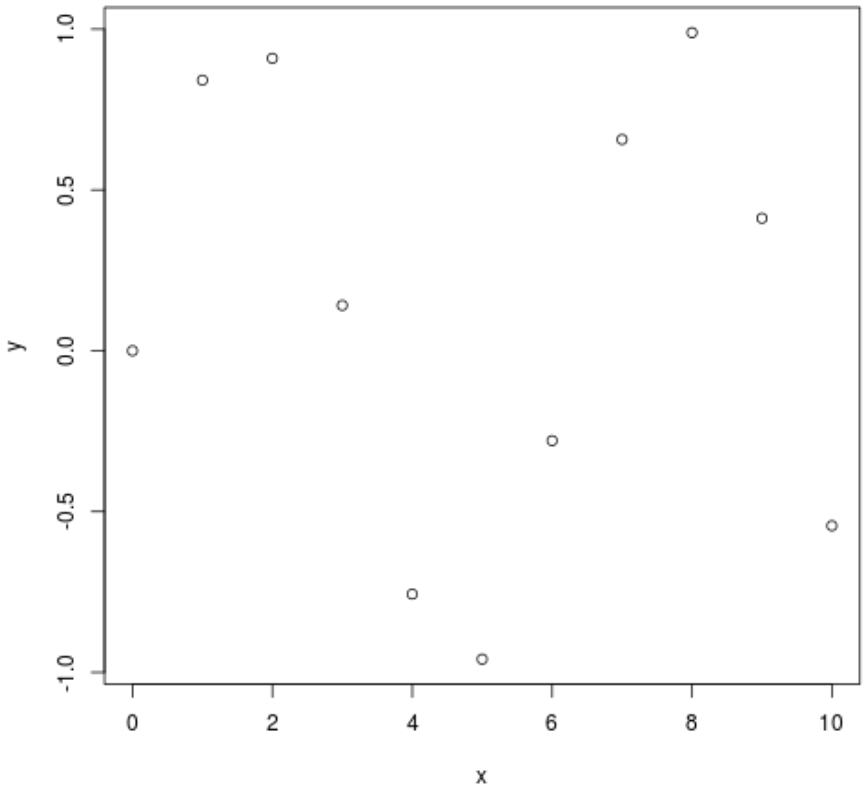
```
x <- rnorm(10);

plot(x)
```



```
plot(x, type="l", col="red", main="Line Diagramm", xlab="time [year]", ylab="differen
```

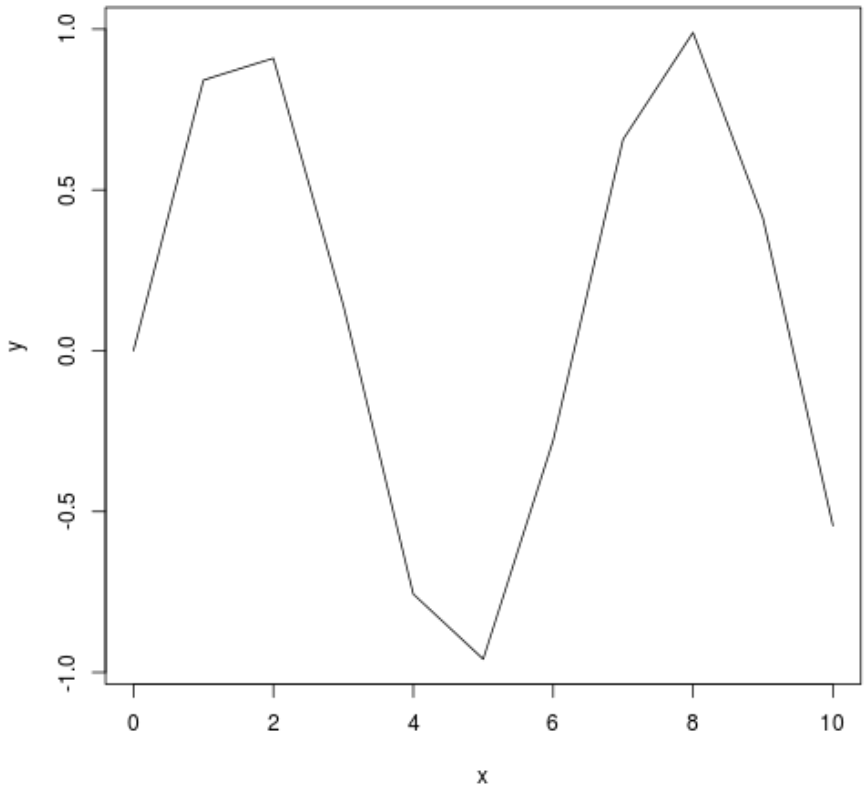



```
x <- (0:10)
y <- sin(x)

x

##  [1]  0  1  2  3  4  5  6  7  8  9 10

y

##  [1]  0.0000000  0.8414710  0.9092974  0.1411200 -0.7568025 -0.9589243
##  [7] -0.2794155  0.6569866  0.9893582  0.4121185 -0.5440211

plot(x, y)
```

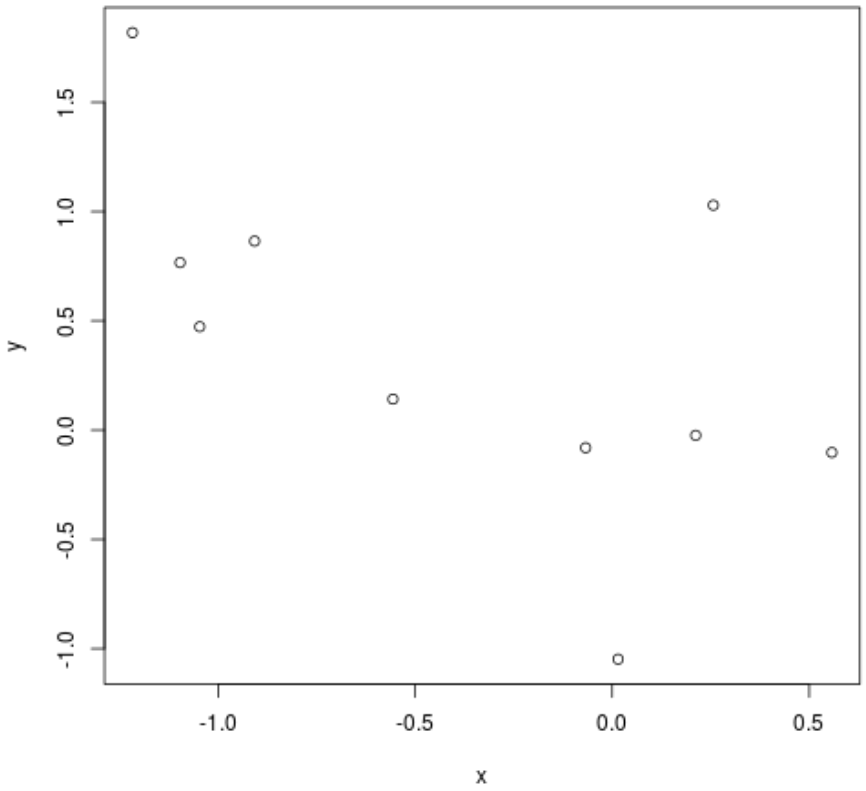



```
plot(x, y, type="l")
```

10

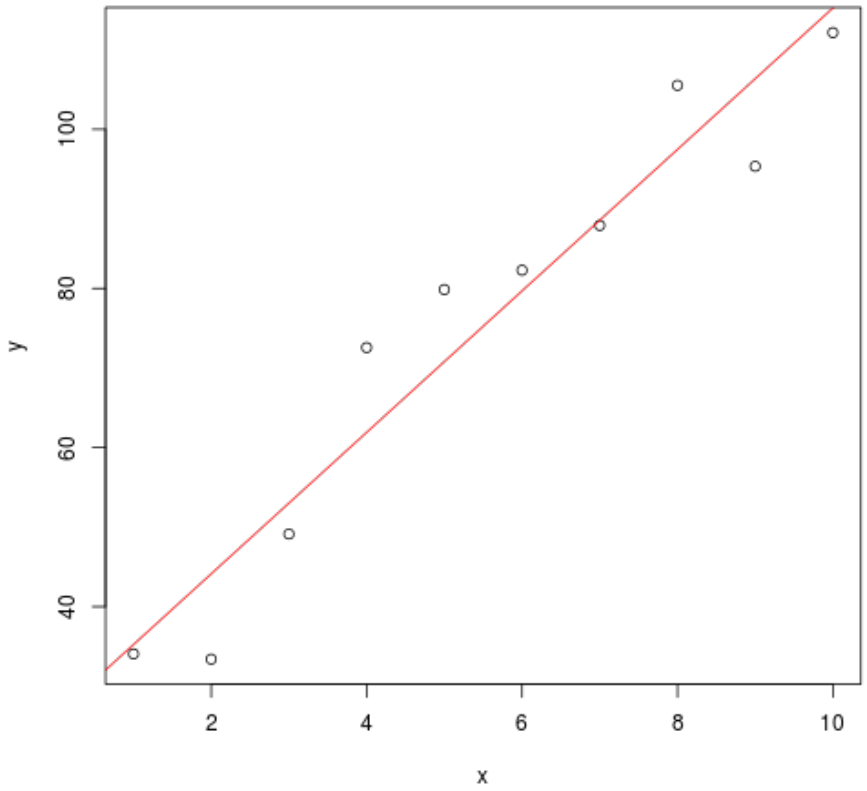



```
x <- rnorm(10); y <- rnorm(10)
plot(x,y)
```




```
x <- 1:10
w <- 20 + 10*x
w

##  [1]  30  40  50  60  70  80  90 100 110 120

linear_sample_df <- data.frame(x=x, y=w + rnorm(10)*10)

plot(linear_sample_df)

linear_model_lm <- lm(y ~ x, data=linear_sample_df)
summary(linear_model_lm)

##
## Call:
## lm(formula = y ~ x, data = linear_sample_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9983  -3.6736  -0.8972   6.7122  10.6759
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.3427     5.5620   4.736  0.00147 **
## x             8.8894     0.8964   9.917 9.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.142 on 8 degrees of freedom
## Multiple R-squared:  0.9248,Adjusted R-squared:  0.9154
## F-statistic: 98.34 on 1 and 8 DF,  p-value: 9.034e-06

abline(linear_model_lm, col="red")
```



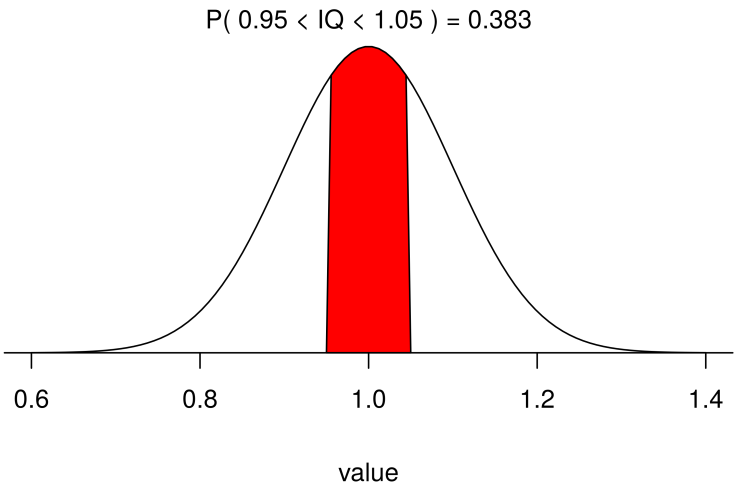


Figure 1: Normal Distribution.

Variance Tests

Significance Tests

Distributions

Normal Distribution

Normal Distribution

Based on the equation

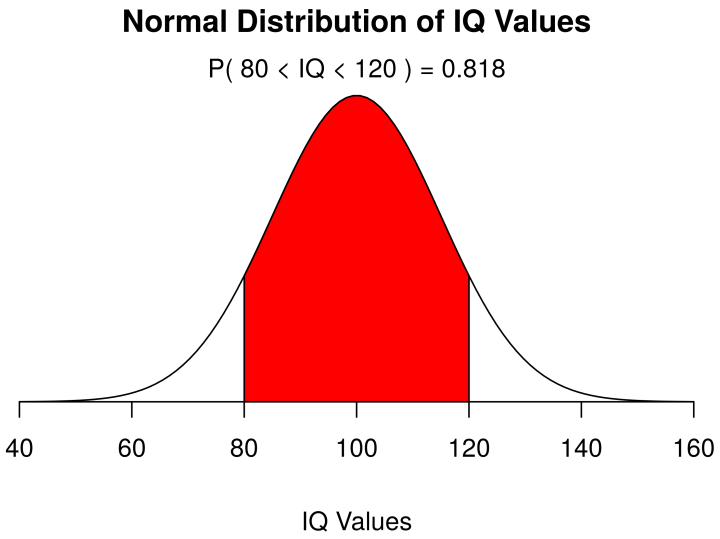
$$f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ = mean σ = standard deviation

with $mean = 1$, $\sigma = 0.1$

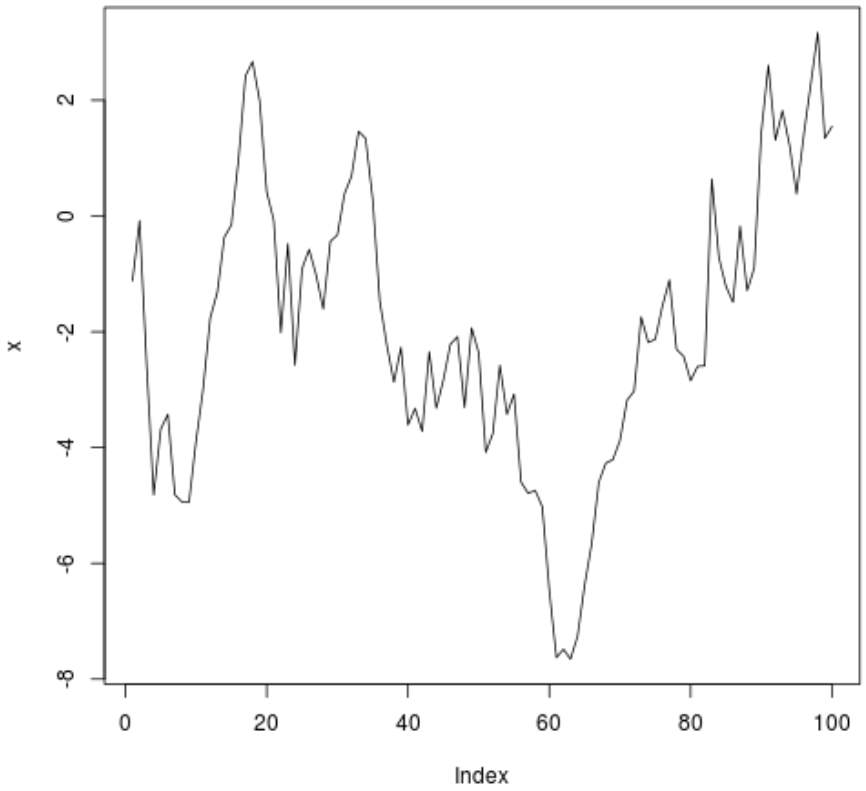
Example of a Normal Distribution

Children's IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. What proportion of children are expected to have an IQ between 80 and 120?



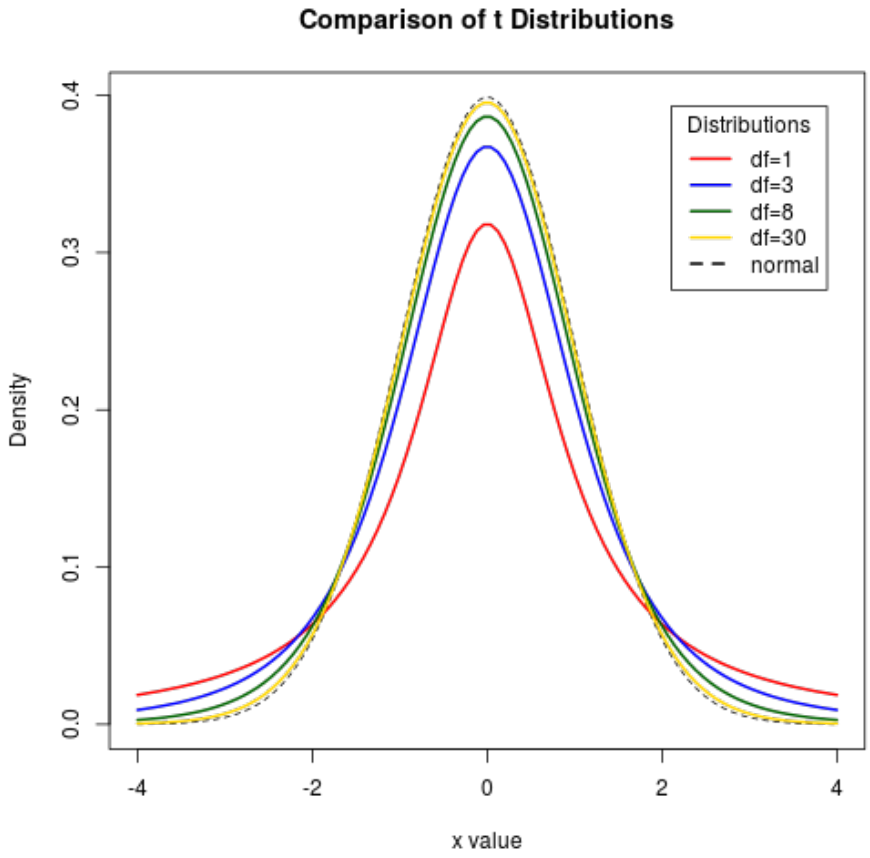
Cusum Example

```
## [1] -1.939758
```



Student Distribution

Display the Student's t distributions with various degrees of freedom and compare to the normal distribution


Comparison of t Distributions
Distributions
df=1
df=3
df=8
df=30
normal
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x value