Statistics in Julia

Julia has been used by mathematicians primarily over its history and therefore has a rich mathematic ecosystem

In [1]:
```
using DataFrames, RDatasets
iris = dataset("datasets", "iris");
```

In [2]:
```
using Statistics, Printf
for n in names(iris)[1:4]
    @printf("%s\t%0.3f\t%0.3f\t%0.3f\n", string(n),
        mean(iris[!,n]), std(iris[!, n]), cov(iris[!, n], iris[!, :Species].refs
))
end
```

```
SepalLength     5.843   0.828   0.531
SepalWidth      3.057   0.436   -0.152
PetalLength     3.758   1.765   1.372
PetalWidth      1.199   0.762   0.597
```

Distributions.jl

Used to generate data according to distributions (as seen in the previous section) or to fit distributions to data

In [3]:
```
using Distributions
```

We can use Maximum Likelihood Estimation to fit a distribution

In [4]:
```
X = iris[!, :SepalLength]
d = fit_mle(Normal, X)
```

Out[4]:
```
Normal{Float64}(μ=5.843333333333335, σ=0.8253012917851409)
```

To compare the result, we can generate data from this distribution and calculate the mean squared error.

In [5]:
```
y = rand(d, length(X));
mean((X .- y).^2)
```

Out[5]:
```
1.2584806167138878
```

GLM.jl

Generalized Linear Models for linear regression. We'll look at ordinary least squares regression

In [6]:
```
using GLM
```

In [7]:
```
iris[!, :Sind] = iris[!, :Species].refs
first(iris, 6)
```

Out[7]:

6 rows × 6 columns

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	Sind
	Float64	Float64	Float64	Float64	Categorical...	UInt8
1	5.1	3.5	1.4	0.2	setosa	0x01
2	4.9	3.0	1.4	0.2	setosa	0x01
3	4.7	3.2	1.3	0.2	setosa	0x01
4	4.6	3.1	1.5	0.2	setosa	0x01
5	5.0	3.6	1.4	0.2	setosa	0x01
6	5.4	3.9	1.7	0.4	setosa	0x01

In [8]:
```
ols = lm(@formula(Sind ~ SepalLength), iris)
```

Out[8]:

```
StatsModels.TableRegressionModel{LinearModel{GLM.LmResp{Array{Float64,1}},GLM.
DensePredChol{Float64,LinearAlgebra.Cholesky{Float64,Array{Float64,2}}}},Array
{Float64,2}}

Sind ~ 1 + SepalLength

Coefficients:
──────────────────────────────────────────────────────────────────────────────
                Estimate  Std. Error   t value  Pr(>|t|)  Lower 95%  Upper 95%
──────────────────────────────────────────────────────────────────────────────
(Intercept)  -2.52398    0.29878    -8.44763    <1e-13  -3.11441   -1.93356
SepalLength   0.774212   0.0506293  15.2918     <1e-31   0.674163   0.874262
──────────────────────────────────────────────────────────────────────────────
```

In [ ]: