MendelGeneticCounseling ASHG Workshop - creating your own functions to calculate risk of disease.

last update: October 12 2019, 4pm.

The purpose of this tutorial is to demonstrate how to use the probabilities of the genotypes calculated in the MendelGeneticCounseling tutorial to calculate the probability of the phenotype. To accomplish this we create a function that can calculate a risk of a phenotype using results of genotype prediction, thus also showing how to write simple functions in Julia.

NOTE: When finished with this notebook, go to the File tab and first select 'save and checkpoint' to save your results. Then under the File tab, select 'close and halt' to prevent copies of Jupyter notebook from running indefinitely in the background.

Check the version of Julia you will be using:

For reproducibility, check the machine information below. To execute a notebook command, hold down`Shift Enter` within the box. This tutorial and corresponding modules have been checked with Julia versions 1.0, 1.1 and 1.2. Please report any issues running the tutorial or the module to Janet Sinsheimer PhD. (jsinshei@g.ucla.edu).

In [1]:
```
versioninfo()
```

```
Julia Version 1.2.0
Commit c6da87ff4b (2019-08-20 00:03 UTC)
Platform Info:
  OS: macOS (x86_64-apple-darwin18.6.0)
  CPU: Intel(R) Core(TM) i7-6567U CPU @ 3.30GHz
  WORD_SIZE: 64
  LIBM: libopenlibm
  LLVM: libLLVM-6.0.1 (ORCJIT, skylake)
```

This tutorial uses the Julia modules: Distributions, LinearAlgebra, CSV, DataFrames and Delimiteed Files. To add them use the package manager type

] add Distributions, LinearAlgebra, CSV, DataFrames and Delimiteed Files

In [2]:
```
using Distributions, LinearAlgebra, CSV, DataFrames, DelimitedFiles
```

Check your working directory. For convenience have the julia notebook in the same directory as your files.

In [3]:
```
pwd()
```

Out[3]:
```
"/Users/janets/Documents/lap_top_janet/janet/short_courses_2019/ASHGproposal/Gen
eticCounseling_ASHG2019-master10062019"
```

In [ ]:
```
#If you need to change directories you can use cd(). The exact syntax depends on w
hether you are using a Mac or a PC (windows).

#For the Mac an example is:  cd("/Users/janets/GeneticCounseling/Chol")

#For the PC, an example is: cd("C:\\Users\\Janet Sinsheimer\\Documents\\Julia_file
s")
```

Example 1: Calculate the risk when a penetrance file is provided

This example is very simple but it serves to introduce the concepts. The example calculates the probability that a woman is affected with Breast Cancer at or before a specified age (in this example before or during her sixties). It uses the probabilities of an individual's genotypes given their covariates and their family's information as calculated by MendelGeneticCounseling.jl. The example uses the same breast cancer data provided in the MendelGeneticCounselingTutorial.iynb. In particular use the cumulative penetrance classes provided for BRCA1 (http://analyzemyvariant.com/brca1-info (http://analyzemyvariant.com/brca1-info))

We thank Brian Shirts for pointing us to the "Analyze My Variant" website (http://analyzemyvariant.com (http://analyzemyvariant.com)) for a realistic example.

BrianspedigreeMay162019.jpg

Step 1: Calculate the probability that individual 18 has each of the three possible underlying genotypes.

Change the pedigree file by changing the genotype for individual 18 in the top pedigree in order to calculate

$P(G_{18} = 1/1|\mathbf{G, CancerStatus, Age, Sex})$,

$P(G_{18} = 1/2|\mathbf{G, CancerStatus, Age, Sex})$,

and

$P(G_{18} = 2/2|\mathbf{G, CancerStatus, Age, Sex})$,

In the example ControlBRCAExample.txt individual 18 is heterozygous in the top pedigree. You will need to change the pedigree file and change the pedigree and outfile names in the control file. Then run MendelGeneticCounseling three times, once with each of the three pedigree files.

In [4]:
```
using MendelGeneticCounseling
GeneticCounseling("PhenotypeRisk/ControlBRCAHZNExample.txt")
```

```
     Welcome to OpenMendel's
 Genetic Counseling Analysis Option


Reading the data.

The current working directory is "/Users/janets/Documents/lap_top_janet/janet/sh
ort_courses_2019/ASHGproposal/GeneticCounseling_ASHG2019-master10062019/Phenotyp
eRisk".

Keywords modified by the user:

  control_file = PhenotypeRisk/ControlBRCAHZNExample.txt
  disease_status = Cancer
  locus_file = LocusBRCAExample.txt
  output_file = BRCAExampleHZNOut.txt
  pedigree_file = PedBRCAHZNExample.csv
  penetrance_file = PenBRCAExample.csv
  phenotype_file = PhenoBRCAExample.txt

this problem has 2 factors called Symbol[:Sex, :Risk_decade]

Analyzing the data.

 The risk =  0.54887.


Mendel's analysis is finished.
```

In [5]:
```
GeneticCounseling("PhenotypeRisk/ControlBRCAExample.txt")
```

```
     Welcome to OpenMendel's
 Genetic Counseling Analysis Option


Reading the data.

The current working directory is "/Users/janets/Documents/lap_top_janet/janet/sh
ort_courses_2019/ASHGproposal/GeneticCounseling_ASHG2019-master10062019/Phenotyp
eRisk".

Keywords modified by the user:

  control_file = PhenotypeRisk/ControlBRCAExample.txt
  disease_status = Cancer
  locus_file = LocusBRCAExample.txt
  output_file = BRCAExampleOut.txt
  pedigree_file = PedBRCAExample.csv
  penetrance_file = PenBRCAExample.csv
  phenotype_file = PhenoBRCAExample.txt

this problem has 2 factors called Symbol[:Sex, :Risk_decade]

Analyzing the data.

 The risk =  0.45091.


Mendel's analysis is finished.
```

In [6]:
```
GeneticCounseling("PhenotypeRisk/ControlBRCAHZMExample.txt")
```

```
     Welcome to OpenMendel's
 Genetic Counseling Analysis Option


Reading the data.

The current working directory is "/Users/janets/Documents/lap_top_janet/janet/sh
ort_courses_2019/ASHGproposal/GeneticCounseling_ASHG2019-master10062019/Phenotyp
eRisk".

Keywords modified by the user:

  control_file = PhenotypeRisk/ControlBRCAHZMExample.txt
  disease_status = Cancer
  locus_file = LocusBRCAExample.txt
  output_file = BRCAExampleOut.txt
  pedigree_file = PedBRCAHZMExample.csv
  penetrance_file = PenBRCAExample.csv
  phenotype_file = PhenoBRCAExample.txt

this problem has 2 factors called Symbol[:Sex, :Risk_decade]

Analyzing the data.

 The risk =  0.00021.


Mendel's analysis is finished.
```

To summarize, we find that

$P(G_{18} = 1/1|\mathbf{G, CancerStatus, Age, Sex}) = 0.54887,$

$P(G_{18} = 1/2|\mathbf{G, CancerStatus, Age, Sex}) = 0.45091,$

and

$P(G_{18} = 2/2|\mathbf{G, CancerStatus, Age, Sex}) =$0.00021.

Step 2: Examine the penetrance file

We have set up the penetrance file to have 5 columns. The first column is for the risks for homozygous wild type genotype (1/1) for each sex and risk decade (penetrance class). The second column is for the risks for heterozyous genotype (1/2) by penetrance class. The third column is for the risks for the homozygous high risk genotype (2/2) by penetrance class. The next column corresponds to the sex of the individual. The final column corresponds to the risk decade of the individual (1: $0 \leq age < 20$, 2: $20 \leq age < 30$, 3: $30 \leq age < 40$, 4: $40 \leq age < 50$, 5: $50 \leq age < 60$, 6: $60 \leq age < 70$, and 7: $70 \leq age$).

Read the file into a data frame and then extract the penetrances for women age 60 or greater but less than 70 into a vector.

In [7]:
```
BRCApen=CSV.read("PhenotypeRisk/PenBRCAExample.csv",;header=1)
```

Out[7]: 14 rows × 5 columns

	Homozygous_Normal	Heterozygous	Homozygous_Mutant	Sex	Risk_decade
	Float64	Float64	Float64	String	Int64
1	8.85e-7	0.0010259	0.0010259	female	1
2	4.0997e-5	0.047524	0.047524	female	2
3	0.00189916	0.18042	0.18042	female	3
4	0.00878848	0.3736	0.3736	female	4
5	0.0275136	0.5752	0.5752	female	5
6	0.05646	0.6889	0.6889	female	6
7	0.0793	0.785	0.785	female	7
8	7.58e-8	1.07e-5	1.07e-5	male	1
9	1.2e-6	0.00017	0.00017	male	2
10	1.9e-5	0.0012	0.0012	male	3
11	8.5e-5	0.003	0.003	male	4
12	0.00027	0.0062	0.0062	male	5
13	0.00067	0.012	0.012	male	6
14	0.0012	0.018	0.018	male	7

In [8]:
```
pen = (BRCApen[6,1],BRCApen[6,2],BRCApen[6,3])
```

Out[8]: (0.05646, 0.6889, 0.6889)

Step 3: Write a function that calculates the risk given the penetrance file, risk factors, and genotype probabilities

Note that once the genotypes of individual 18 is specified the cumulative penetrance doesn't depend on their families genotypes or phenotypes. This function is quite simple but it illustrates how to write a function.

In [9]:
```
function risknon(probgeno::Tuple,data::Tuple)where T<:Float64
        probpheno=probgeno[1]*data[1]+probgeno[2]*data[2]+probgeno[3]*data[3]
        return probpheno
end
risknon #uses the cumulative penetrances and the conditional probabilies of the ge
notypes
```

Out[9]: risknon (generic function with 1 method)

Genotype probabilities:

In [10]:
```
prob1 = 0.54887
prob2 = 0.45091
prob3 = 0.00021
probgeno = (prob1, prob2, prob3)
```

Out[10]: (0.54887, 0.45091, 0.00021)

In [11]:
```
risknon(probgeno, pen)
```

Out[11]: 0.3417657682

Conclusion

We find that individual 18 has a probability of approximately 34% of developing breast cancer on or before her sixties.

Example 2: Calculate the risk that individual IVII will have a cholesterol value of 300 or greater at age 20.

Data used in Example 2:

The input files for all examples in this tutorial can be obtained from https://github.com/OpenMendel/GeneticCounseling_ASHG2019 (https://github.com/OpenMendel/GeneticCounseling_ASHG2019)

The data are from an example pedigree used in the Mendel version 16.0 release. The pedigree structure and phenotypes are originally from Schrott et al. (1972) Ann Int Med 76:711–720. We have used the pedigree to provide a slightly contrived example in which Mother III13, who has cholesterol value 440 at age 21 is concerned that her young son might also be affected with extreme hypercholesterolemia.

Step 1: Calculate the probability of each of the possible genotypes using MendelGeneticCounseling

This tutorial assumes you have already worked with the MendelGeneticCounseling module previously. If not, please the see the tutorial MendelGeneticCounselingtutorial.ipynb.

Run MendelGeneticCounseling three times.

Using your favorite editor, alter the pedigree file by changing person IV11's genotype in the top pedigree. Also change the control file to use the new pedigree file and save the result in a new file. Then run MendelGeneticCounseling three times. The first time, determine the probability that individual IV11 has genotype -/-, the second time determine the probability that individual IV11 ahs genotype -/+, and the third time determine the probability that individual IV11 has genotype +/+

In [12]:
```
# first call of the GeneticCounseling function takes long because of JIT compiling
# homozygous wild type genotype
# if running only this example remove the `#` from the next line
# using MendelGeneticCounseling
GeneticCounseling("PhenotypeRisk/ControlParametricPenetranceHZNExample.txt")
```

```
     Welcome to OpenMendel's
 Genetic Counseling Analysis Option


Reading the data.

The current working directory is "/Users/janets/Documents/lap_top_janet/janet/sh
ort_courses_2019/ASHGproposal/GeneticCounseling_ASHG2019-master10062019/Phenotyp
eRisk".

Keywords modified by the user:

  control_file = PhenotypeRisk/ControlParametricPenetranceHZNExample.txt
  glm_link = LogLink
  glm_mean = 4.691+0.562(max(allele1,allele2))+0.00194Age+0.036Sex
  glm_response = GammaDist
  glm_scale = 44.68
  glm_trait = Chol
  glm_trials = 1
  locus_file = LocusChol.txt
  output_file = CholHomozygous_NormalRisk.txt
  pedigree_file = PedCholHZN.csv
  phenotype_file = PhenoChol.txt

 no penetrance file

Analyzing the data.

 The risk =  0.71335.


Mendel's analysis is finished.
```

In [13]:
```
# heterozygous genotype
GeneticCounseling("PhenotypeRisk/ControlParametricPenetranceExample.txt")
```

```
     Welcome to OpenMendel's
 Genetic Counseling Analysis Option


Reading the data.

The current working directory is "/Users/janets/Documents/lap_top_janet/janet/sh
ort_courses_2019/ASHGproposal/GeneticCounseling_ASHG2019-master10062019/Phenotyp
eRisk".

Keywords modified by the user:

  control_file = PhenotypeRisk/ControlParametricPenetranceExample.txt
  glm_link = LogLink
  glm_mean = 4.691+0.562(max(allele1,allele2))+0.00194Age+0.036Sex
  glm_response = GammaDist
  glm_scale = 44.68
  glm_trait = Chol
  glm_trials = 1
  locus_file = LocusChol.txt
  output_file = CholHeterozygousRisk.txt
  pedigree_file = PedChol.csv
  phenotype_file = PhenoChol.txt

 no penetrance file

Analyzing the data.

 The risk =  0.27557.


Mendel's analysis is finished.
```

In [14]:
```
# homozygous mutant genotype
GeneticCounseling("PhenotypeRisk/ControlParametricPenetranceHZMExample.txt")
```

```
     Welcome to OpenMendel's
 Genetic Counseling Analysis Option


Reading the data.

The current working directory is "/Users/janets/Documents/lap_top_janet/janet/sh
ort_courses_2019/ASHGproposal/GeneticCounseling_ASHG2019-master10062019/Phenotyp
eRisk".

Keywords modified by the user:

  control_file = PhenotypeRisk/ControlParametricPenetranceHZMExample.txt
  glm_link = LogLink
  glm_mean = 4.691+0.562(max(allele1,allele2))+0.00194Age+0.036Sex
  glm_response = GammaDist
  glm_scale = 44.68
  glm_trait = Chol
  glm_trials = 1
  locus_file = LocusChol.txt
  output_file = CholHomozygous_MutantRisk.txt
  pedigree_file = PedCholHZM.csv
  phenotype_file = PhenoChol.txt

 no penetrance file

Analyzing the data.

 The risk =  0.01108.


Mendel's analysis is finished.
```

Summarizing, we have calculated that:

$P(G_{IV11} = -/- | \mathbf{G, Chol, Age, Sex}) = 0.71335$

$P(G_{IV11} = -/+ | \mathbf{G, Chol, Age, Sex}) = 0.27557$

and

$P(G_{IV11} = +/+ | \mathbf{G, Chol, Age, Sex}) = 0.01108$

We will use these values to calculate the risk that individual IV11 will have a cholesterol value greater than 300 at age 20. Note that in this simple example, once we have these estimates of for the genotypes, the family members' data are no longer needed to calculate the risk.

Step 2: Create a simple Julia function to calculate the risk by summing over the penetrances.

This function calculates the probability that an individual will have a phenotype value greater than a specified value of X (the threshold) given their known risk factors and the probability of underlying genotypes. Note as written here this function is specific to gamma distributed traits.

In [15]:
```
function riskgamma(A::T,S::T,X::T,sex::T,probgeno::Tuple,coef::Tuple,geno::Tuple)w
here T<:Float64
         mean_HZN = exp(coef[1]+coef[2]*geno[1]+coef[3]*A+coef[4]*sex)/S
         mean_HZM = exp(coef[1]+coef[2]*geno[2]+coef[3]*A+coef[4]*sex)/S
         mean_Het = exp(coef[1]+coef[2]*geno[3]+coef[3]*A+coef[4]*sex)/S
         p_HZN=1.0-cdf.(Gamma(S,mean_HZN),X)
         p_HZM=1.0-cdf.(Gamma(S,mean_HZM),X)
         p_Het = 1.0-cdf.(Gamma(S,mean_Het),X)
         probrisk=probgeno[1]*p_HZN+probgeno[2]*p_Het+probgeno[3]*p_HZM
         return probrisk
  end
riskgamma #uses cdf to calculate risk of a value higher than stated value for the
individual
# if the relevant risk is when a value falls below a threhold then change 1-cdf. t
o cdf.
# if an interval is relevant, e.g. X_1 to X_2 then use cdf.(...,X_2) - cdf.(..., X
_1)
# modify mean function to fit the number of environmental covariates.
# change distribution to suit application
```

Out[15]:
```
riskgamma (generic function with 1 method)
```

Step 3: Specify the input information and run the function.

In [16]:
```
Age = 20.0
sex = 1.0
Threshold = 300.0

prob1=0.71335
prob2 = 0.27557
prob3 = 0.01108
probgeno = (prob1,prob2,prob3)

geno = (1.0, 2.0, 2.0)
coef = (4.691,0.562,0.00194,0.036)
Scale = 44.68
```

Out[16]:
```
44.68
```

In [17]:
```
risk = riskgamma(Age,Scale,Threshold,sex,probgeno,coef,geno)
```

Out[17]:
```
0.2533785208203602
```

Conclusion

We find that individual IV11 has an approximate 25% probability of having a total cholesterol level greater than 300 at age 20.

This very simple function can be made more general with more covariates or a user specified distribution but it gives you an idea of how easy it is to program in Julia.

Final Comments

The Julia version of Mendel, provides an opportunity for the user to easily modify the code to suit their own needs. All the source code is provided and Julia is both accessible and very fast.

Reference

For publication please cite:

OPENMENDEL: a cooperative programming project for statistical genetics. Zhou H, Sinsheimer JS, Bates DM, Chu BB, German CA, Ji SS, Keys KL, Kim J, Ko S, Mosher GD, Papp JC, Sobel EM, Zhai J, Zhou JJ, Lange K. Hum Genet. 2019 Mar 26. doi: 10.1007/s00439-019-02001-z

NOTE: When Finished with this notebook. Go to the File tab and first select save and checkpoint to save your results. Then under the file tab, select close and halt to prevent copies of Jupyter notebook from running indefinitely in the background