Homework

Statistics warm-up

Consider a random variable X with a probability density function

$$f(x) = \frac{c}{\sqrt{x(b-x)}}, \quad 0 < x < b,$$

where c is a normalising constant and b is a parameter.

- Find c such that the probability density function is valid.

X new = (X – X min) / (X max – X min) c=(x-0)/(b-0) c=x/b

- Find the corresponding cumulative distribution function, $F(x)$ and its inverse $F^{-1}(x)$.

CDF = integral(f(x)) from -infinity to x

$$f(x) = \frac{x/b}{\sqrt{x(b-x)}}, \quad 0 < x < b,$$

$$F(x) = \int_{-inf}^{x} \frac{x/b}{\sqrt{x(b-x)}} dx$$

$$x = \int_{-inf}^{F^1(x)} \frac{F^1(x)/b}{\sqrt{F^1(x)(b-F^1(x))}} dx$$

- Describe a procedure to generate samples distributed according to X, given a set of uniformly distributed samples $u_1, \ldots, u_n \sim \mathcal{U}[0,1]$.
- transform uniformily distributed samples so that they are distributed according to X
- use above function (f(x)) to do this
- Set the seed to 42 and generate 1,000 samples distributed according to X using 1,000 uniformly distributed i.i.d. $\mathcal{U}[0,1]$ samples, with $b = 5$. Plot a histogram of the resulting samples.

$$f(x) = \frac{x}{5\sqrt{x(5-x)}}, \quad 0 < x < 5$$

```
set.seed(42)
b<-5
s<-sample(X,1000)
```

```
## Error in sample(X, 1000): object 'X' not found
```

```
hist(s)
```

```
## Error in hist(s): object 's' not found
```

- Assume that we have a set of samples $X_1, \ldots, X_n$ which are i.i.d. as X with b unknown. Find the log-likelihood function for the parameter b, $\ell(b; x_1, \ldots, x_n)$ and find its derivative with respect to b, $\frac{d\ell}{db}$.
- Based on the likelihood, or otherwise, give a statistic based on $X_1, \ldots, X_n$ which is a reasonable estimator for the unknown parameter b. Explain your choice.
- The file `samples.rds` contains samples $X_1, \ldots, X_n$ distributed according to X. Report the estimated value of $\hat{b}$ based on those samples.

$$f(x) = \frac{x/b}{\sqrt{x(b-x)}}, \quad 0 < x < b,$$

```
summary(samples)
```

```
## Error in summary(samples): object 'samples' not found
```

```
t.test(samples)
```

```
## Error in t.test(samples): object 'samples' not found
```

Arrays

You are given a multi-dimensional array `array.rds`. It contains penetrance curves for various cancers and genes. In simple terms, penetrances are how likely one will develop a cancer given that they have a certain corresponding gene mutation. Other variables in `data` describe different sub-populations. For example, the probabilities corresponding to `Brain` cancer and gene `APC` are the probabilities for which a person will develop brain cancer, given that they have a mutation in the `APC` gene.

- Read in the data and give the dimensions of the array.

```
str(array)
```

```
## function (data = NA, dim = length(data), dimnames = NULL)
```

array dimensions: [1:18, 1:26, 1:8, 1:2, 1:94]

Hint: Use the ***`str`*** *function.*

- Subset the array for the penetrances associated with `Breast` cancer and the `BRCA2` gene for a female with the default race `All_Races`. Then plot the penetrance curve (probability versus age).

```
subarray1 <- array[c("Breast"),c("BRCA2"),c("All_Races"),c("Female"), ]
```

```
## Error in array[c("Breast"), c("BRCA2"), c("All_Races"), c("Female"), ]: object of type 'closure' is n
```

```
plot(subarray1,
     main="Probability of Breast cancer given BRCA2 vs. Age",
     xlab="Age",
     ylab="Probability of Breast cancer given BRCA2")
```

```
## Error in plot(subarray1, main = "Probability of Breast cancer given BRCA2 vs. Age", : object 'subarra
```

- Subset the array for the penetrances associated with `Colorectal` cancer and the `PALB2` gene for an Asian male. What is the probability that a person from this subpopulation at age 50 will develop colorectal cancer in the next 10 years given that he has tested positive for a PALB2 mutation but is otherwise disease free?

Hint: The probability over a period of time is calculated by summing the yearly risks.

```
subarray2 <- array[c("Colorectal"),c("PALB2"),c("Asian"),c("Male"),c(50:60) ]
```

```
## Error in array[c("Colorectal"), c("PALB2"), c("Asian"), c("Male"), c(50:60)]: object of type 'closure
```

```
sum(subarray2)
```

```
## Error in eval(expr, envir, enclos): object 'subarray2' not found
```

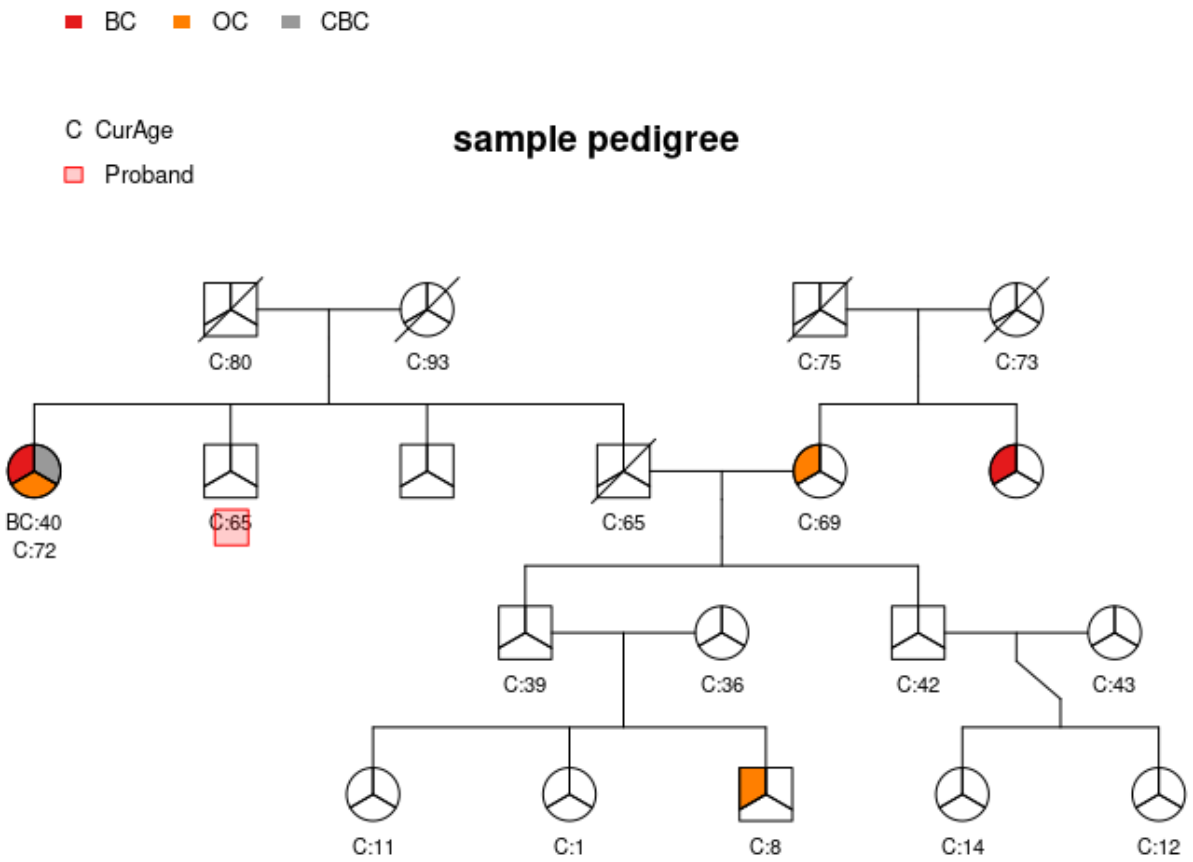

Figure 1:

Family pedigrees

Read in the `.rdata` file `pedigree.rda`. Each `data.frame` represents a family. Each individual is uniquely identified by the first column called `ID`. Their sex is coded as `0` for females and `1` for males. Individuals' mother and father are indicated in the `MotherID` and `FatherID` columns. A value of `NA` in these columns means that this person is a so-called 'founder' or that a certain parent is missing.

Each pedigree can be thought of as a family tree. For example, a visualisation of a sample pedigree is shown below. The colours indicate affliction status for cancers as labelled in the legend.

In the following exercises, you are encouraged to modularise and comment on your code.

- Write an R function(s) to count the number of unique nuclear families there are in a certain pedigree. A nuclear family is defined as the set of two parents and all of their children.

count for every unique combination of mother and father ID

Thought Process: remove NA's make unique id for each nuclear family combine mother/father ids to group families create something new (array?) that combines mother/father id–> column count unique ids (table function) now just count the number of rows from above

use function on each family and then sum start of the function:

```
count.nfam <- function(family) {
  removed <- subset(family, !is.na(MotherID) & !is.na(FatherID))
}
```

- Report the number of nuclear families for the pedigrees contained in the `.rda` file.

for fam10 –> 4

if function above worked/was complete:

```
x10<-family(fam10)
```

```
## Error in family(fam10): object 'fam10' not found
```

```
x100<-family(fam100)
```

```
## Error in family(fam100): object 'fam100' not found
```

```
x50<-family(fam50)
```

```
## Error in family(fam50): object 'fam50' not found
```

```
x75<-family(fam75)
```

```
## Error in family(fam75): object 'fam75' not found
```

```
x10+x100+x50+x75
```

```
## Error in eval(expr, envir, enclos): object 'x10' not found
```