Homework_Answer

Chichun Tan

1/19/2021

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.

Answer: To make the pdf above vaild, we want the integral of $f(x)$ with respect to the interval $0 < x < b$ to be 1. That is,

$$\int_0^b f(x)dx = c\int_0^b \frac{1}{\sqrt{x(b-x)}}dx = 1$$

To solve this equation, let $x = t^2$ and so that $dx = 2tdt$, and then we have

$$\begin{aligned}
\int_0^b \frac{1}{\sqrt{x(b-x)}}dx &= \int_0^{\sqrt{b}} \frac{1}{t}\frac{1}{\sqrt{b-t^2}}2tdt \\
&= 2\int_0^{\sqrt{b}} \frac{1}{\sqrt{b-t^2}}dt \\
&= 2\int_0^{\sqrt{b}} \frac{1}{\sqrt{b}}\frac{1}{\sqrt{1-\frac{t}{\sqrt{b}}^2}}dt \\
&= 2[\arcsin(\frac{t}{\sqrt{b}})]_0^{\sqrt{b}} \\
&= \pi
\end{aligned}$$

Therefore,to make the equation above to be 1, $c = \frac{1}{\pi}$.

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

Answer: From the last problem, the density function of x in $0 < x < b$ is $f(x) = \frac{1}{\pi\sqrt{x(b-x)}}$. Therefore, the cdf in $0 < x < b$ becomes

$$F(x) = \int_0^x \frac{1}{\pi\sqrt{u(b-u)}}du = \frac{2}{\pi}\arcsin(\sqrt{\frac{x}{b}})$$

Therefore, the cdf for x is

$$F(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \frac{2}{\pi}\arcsin(\sqrt{\frac{x}{b}}) & \text{if } 0 < x < b \\ 1 & \text{if } x \ge b \end{cases}$$

To find out the inverse function of cdf $F^{-1}(x)$ in $0 < x < b$, assume that $F(x) = u$ where $u \in [0, 1]$, then

$$\begin{aligned}\frac{2}{\pi}\arcsin(\sqrt{\frac{x}{b}}) = u &\Longrightarrow \sqrt{\frac{x}{b}} = \sin(\frac{\pi u}{2}) \\ &\Longrightarrow x = b\sin^2(\frac{\pi u}{2}) \\ &\Longrightarrow F^{-1}(u) = b\sin^2(\frac{\pi u}{2})\end{aligned}$$

Change the variable in the expression and then we have $F^{-1}(x) = b\sin^2(\frac{\pi x}{2})$. Notice that the inverse only exists on $(0, b)$.

- Describe a procedure to generate samples distributed according to X, given a set of uniformly distributed samples $u_1, \ldots, u_n \sim \mathcal{U}[0, 1]$.

Answer: given a set of random sample from $[0, 1]$, we can regard them as a set of cdf values from above function. Therefore, we can generate samples distributed according to X by taking the inverse value of the set.

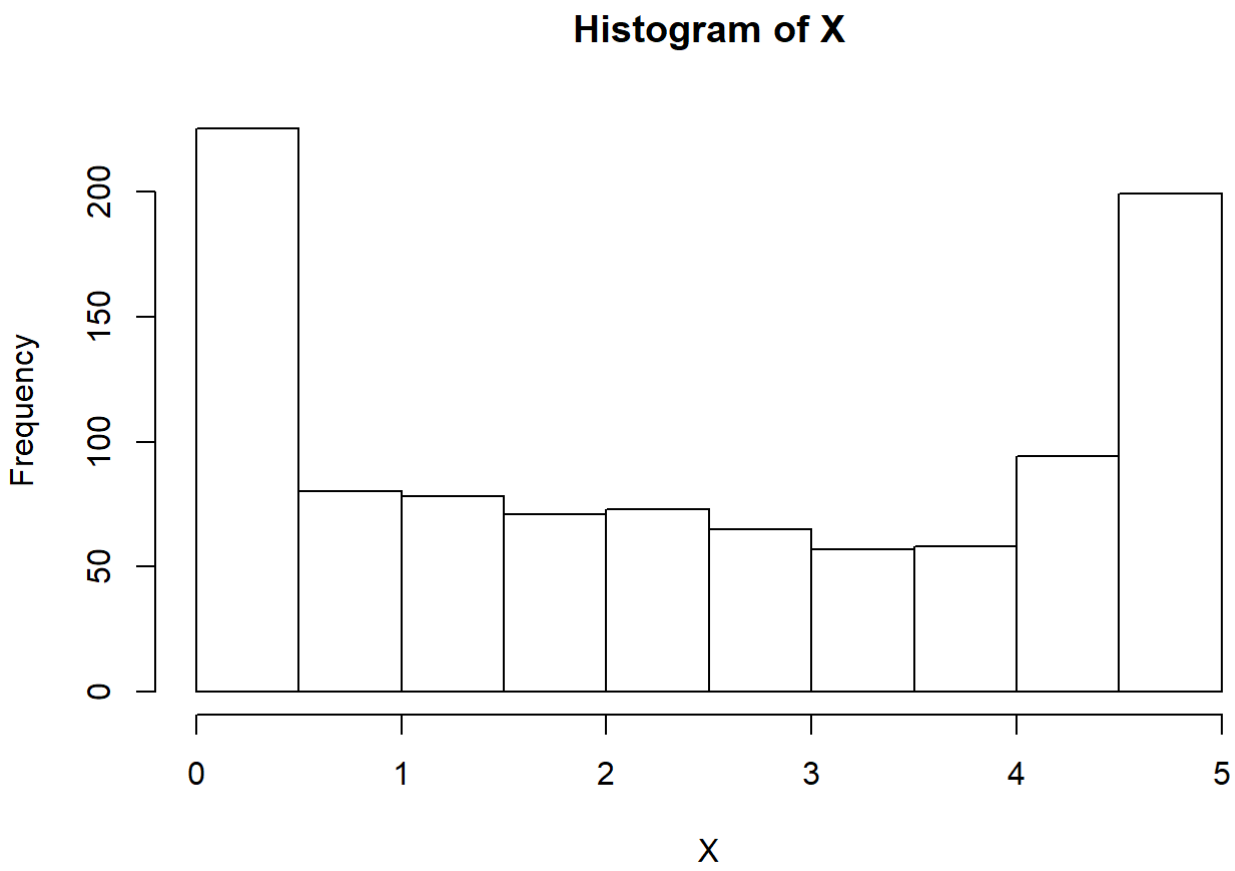
- Set the seed to 42 and generate 1,000 samples distributed according to X using 1,000 uniformly distributed i.i.d. $\mathcal{U}[0, 1]$ samples, with $b = 5$. Plot a histogram of the resulting samples.

```
F_inv <- function(x,b){
  return(b*sin(pi*x/2)^2)
}
```

```
set.seed(42)
b <- 5
F <- runif(1000,0,1)
X <- rep(NA,1000)

for(i in 1:length(F)){
  X[i] <- F_inv(F[i],b)
}

hist(X, xlab = "X")
```



- 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, $\frac{d\ell}{db}$.

Answer:

$$L(b|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\pi\sqrt{x_i(b-x_i)}} = \pi^{-n}(\prod_{i=1}^{n} x_i(b-x_i))^{-\frac{1}{2}}$$

$$logL = \ell(b|x_1, \ldots, x_n) = -n\log\pi - \frac{1}{2}(\sum_{i=1}^{n}\log x_i + \sum_{i=1}^{n}\log(b-x_i))$$

$$\frac{d\ell}{db} = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{b-x_i}$$

* 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.

Answer: Consider the derivative above, because $b \geq \max x_i$, the first derivative is negative. Consider the second derivative,

$$\frac{d^2\ell}{db^2} = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{(b-x_i)^2}$$

The second derivative is positive so that the first derivative is monotonic on b. In other words, the first derivative is constantly negative on b. Therefore, when $b = \max x_i$ we have the maximum likelihood. In conclusion, we have a reasonable estimator $\hat{b} = \max x_i$, which is a maximum likelihood estimator in this case.

- 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.

```
X <- readRDS("data/samples.rds")
b.hat <- max(X)
b.hat
```

```
## [1] 85.19952
```

$\hat{b} = 85.2$ is a reasonable estimator.

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.

Hint: Use the `str` *function.*

Answer:

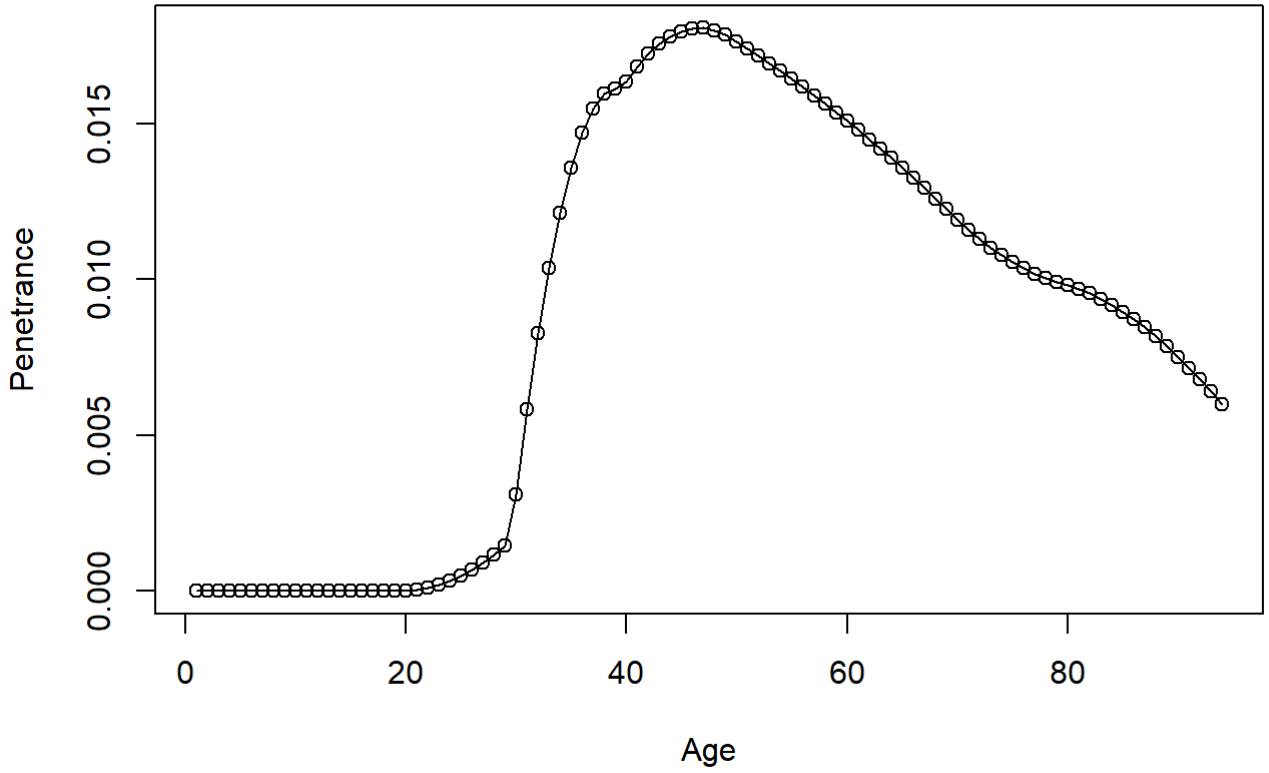
```
data.array <- readRDS("data/array.rds")
str(data.array)
```

```
##  num [1:18, 1:26, 1:8, 1:2, 1:94] 3.98e-05 2.80e-07 0.00 5.00e-08 0.00 ...
##  - attr(*, "dimnames")=List of 5
##   ..$ Cancer: chr [1:18] "Brain" "Breast" "Cervical" "Colorectal" ...
##   ..$ Gene  : chr [1:26] "APC" "ATM" "BARD1" "BMPR1A" ...
##   ..$ Race  : chr [1:8] "All_Races" "AIAN" "Asian" "Black" ...
##   ..$ Sex   : chr [1:2] "Female" "Male"
##   ..$ Age   : chr [1:94] "1" "2" "3" "4" ...
```

For each sub-population level- each race, each sex and each age $(8 \times 2 \times 94)$ - we have a 18×26 matrix for penetrances. In total, the dimension of the array should be $18 \times 26 \times 8 \times 2 \times 94$

- 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).

```
Sub.data1 <- data.array["Breast","BRCA2","All_Races","Female",]
plot(x = 1:length(Sub.data1),y = Sub.data1, xlab = "Age", ylab = "Penetrance", type = "l")
points(x = 1:length(Sub.data1),y = Sub.data1)
```



- 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.

```
Sub.data2 <- data.array["Colorectal","PALB2","Asian","Male",]
risk <- sum(Sub.data2[as.character(50:(50+9))])
risk
```

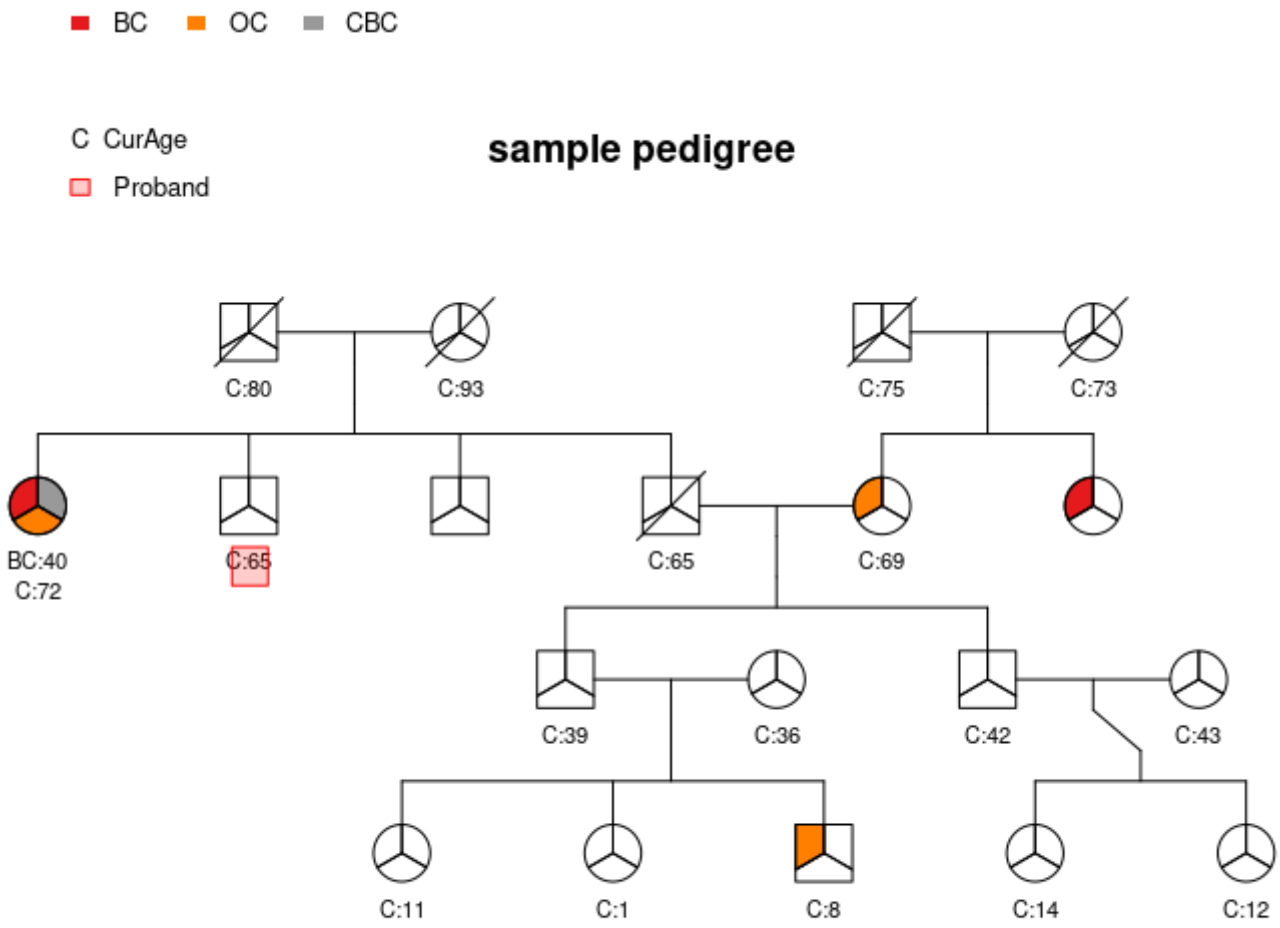
```
## [1] 0.0071847
```

The risk of colorectal cancer in the next 10 years is 0.0072.

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.

```
load("data/pedigree.rda")
```

```
num_nucfam <-function(pedigree){
  df <- cbind(pedigree$MotherID,pedigree$FatherID) # extract parents ID
  df <- df[complete.cases(df),] # remove the subjects with NA parents ID
  return(nrow(unique(df))) # count the number of unique pairs of parents
}
```

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

```
fam10_nucfam <- num_nucfam(fam10)
fam50_nucfam <- num_nucfam(fam50)
fam75_nucfam <- num_nucfam(fam75)
fam100_nucfam <- num_nucfam(fam100)

tb <- matrix(c(fam10_nucfam,fam50_nucfam,fam75_nucfam,fam100_nucfam), ncol = 4)
colnames(tb) <- c("fam10_nucfam","fam50_nucfam","fam75_nucfam","fam100_nucfam")
tb
```

```
##      fam10_nucfam fam50_nucfam fam75_nucfam fam100_nucfam
## [1,]            4           10           18            26
```