EDA A4 Q2

Data

In this stage, the plan is executed. Instead of 1500 records of treatment `A` and `B`, 1600 of each were found. The number of males and females was kept equal (now 1600 of each sex).

The process was to search the records in order, selecting those first encountered to get 1600 for each treatment and 1600 of each sex. Many records might be discarded whenever one quota was met and the search continued to meet the other quotas. It was also noticed that the patient's age was available for each record, so that the effect of treatment on younger and older adults might also be considered.

The counts which fell into the various categories were assembled into the following data set.

```
#file <- path_concat(dataDirectory, "medicalRecords.rda")
#load(file)
```

Alternatively we could have read the data in using the ".csv" file.

```
#file <- path_concat(dataDirectory, "medicalRecords.csv")
medicalRecords <- read.csv('/Users/rudranibhadra/Downloads/medicalRecords.csv')
```

In either case, the data looks like

```
medicalRecords
```

```
##      Age    Sex Treatment   Outcome Freq
## 1  20-39   Male         A Recovered   60
## 2  40-59   Male         A Recovered   90
## 3  20-39 Female         A Recovered  270
## 4  40-59 Female         A Recovered  540
## 5  20-39   Male         B Recovered  450
## 6  40-59   Male         B Recovered   60
## 7  20-39 Female         B Recovered  240
## 8  40-59 Female         B Recovered   50
## 9  20-39   Male         A      Died   40
## 10 40-59   Male         A      Died  210
## 11 20-39 Female         A      Died   30
## 12 40-59 Female         A      Died  360
## 13 20-39   Male         B      Died  450
## 14 40-59   Male         B      Died  240
## 15 20-39 Female         B      Died   60
## 16 40-59 Female         B      Died   50
```

```
summary(medicalRecords)
```

```
##     Age        Sex    Treatment      Outcome       Freq
##  20-39:8   Female:8   A:8        Died     :8   Min.   : 30.0
##  40-59:8   Male  :8   B:8        Recovered:8   1st Qu.: 57.5
##                                                Median :150.0
##                                                Mean   :200.0
##                                                3rd Qu.:292.5
##                                                Max.   :540.0
```

which is turned into a multi-way table of counts using `xtabs()` as follows

```
medicalRecordsTable <- xtabs(Freq ~ Age + Sex + Treatment + Outcome,
                             data = medicalRecords)
#medicalRecordsTable[,,,]
```

This is a multi-way contingency table of dimension

```
dim(medicalRecordsTable)
```

```
## [1] 2 2 2 2
```

of which, by design, the marginal table of sex distribution is

```
margin.table(medicalRecordsTable, margin = c(2))
```

```
## Sex
## Female   Male
##   1600   1600
```

and also of which, by design, the marginal table of treatment distribution is

```
margin.table(medicalRecordsTable, margin = c(3))
```

```
## Treatment
##    A    B
## 1600 1600
```

a total of 3200 records selected from those available.

a. The **Problem** stage.

b. *(2 marks)* What is the target population here? What are the units making up the population?

The target population is the medical records from another country of those who had contracted the disease and had been treated with one of the two treatments. The units are each of the medical records.

ii. *(2 marks)* What are the variates? Which are response variate(s)? Which are explanatory?

The variates are age, sex, treatment, outcome, freq. The explanatory variables are age, sex, treatment and freq. The response variable is outcome.

iii. *(2 marks)* What population attribute might be of interest?

The proportion of people who recovered from each of the treatments.

b. The **Plan** stage.

c. *(3 marks)* What is the study population? How might study error arise?

The study population consists of the Electronic medical records available from several of the more populous districts are accessible. The study error might arise from the fact that the study population may not fully represent the medical records from another country.

ii. *(5 marks)* What is the sampling plan? How might sampling bias arise?

A sample size of n=3000 will be taken and from there 1500 were treated with A and 1500 with B. The number of males and females are equal.

Sampling bias might arise as people under 20 years are not represented in the data set even though they are also more vulnerable to the disease.

iii. *(2 marks)* Is this study experimental or observational? Explain your reasoning.

The study is observational as the records of the medical treatments done on the patients using the two treatment methods were searched from another district and no experiments were done on the patients while getting the data.

c. The **Data** stage.

d. *(2 marks)* What is the sample? How might sample error arise?

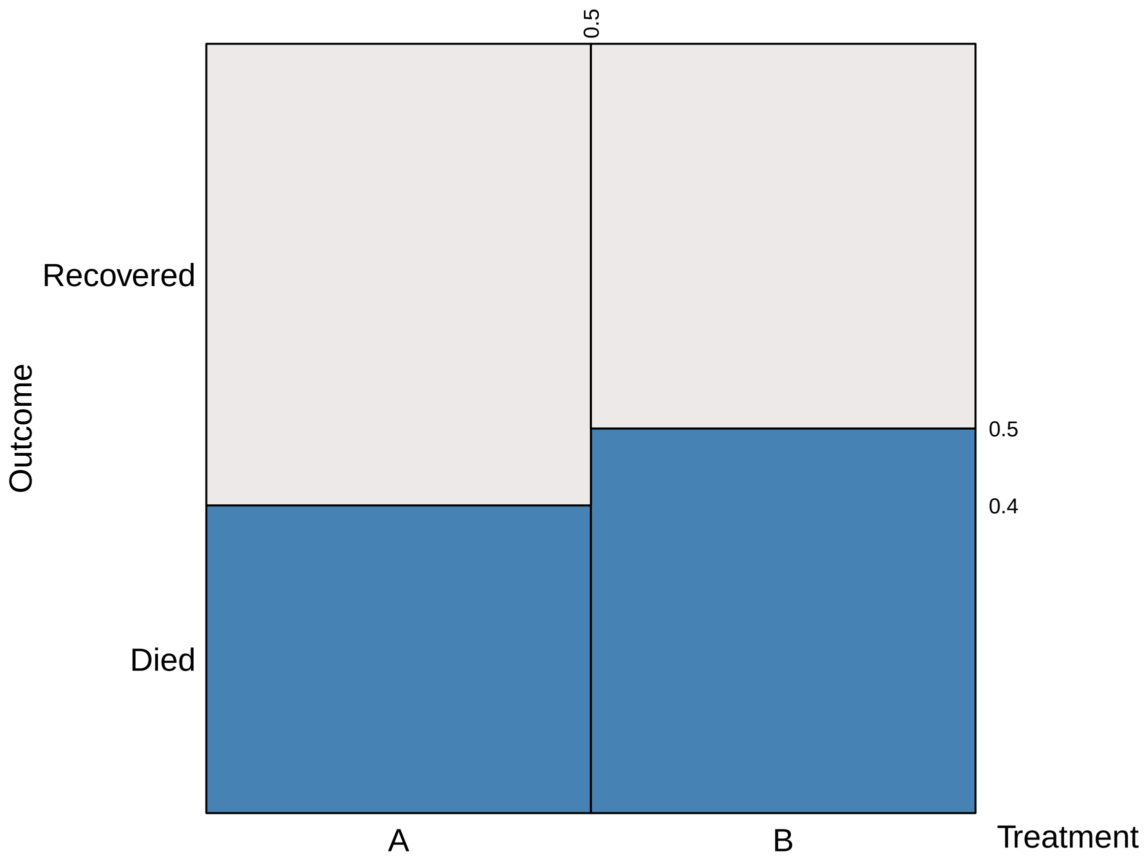
The sample consists of 1600 records of each treatment A and B. Sample error might arise from the fact that the sample might not represent the all the electronic medical records available.

d. The **Analysis** stage

e. *(3 marks)* Construct the eikosogram of `Outcome` versus `Treatment` with title "All patients".

On the basis of this eikosogram, which treatment would be preferred? Why?

```r
library(eikosograms)
eikos(Outcome~Treatment, data = medicalRecordsTable)
```



From the eikosogram, treatment A would be prefered as the outcome to be recovered is more likely when it is treatment A as compared to when it is treatment B.

ii. *(3 marks)* As a sanity check use `chisq.test()` on the marginal table of only `Outcome` versus `Treatment`.

- Show your code for constructing the table and conducting the test.
- What do you conclude about the relationship between treatment and outcome? Why?
- Which treatment would you recommend?

```
t<-margin.table(medicalRecordsTable,margin=c(3,4))
t
```

```
##          Outcome
## Treatment Died Recovered
##         A  640       960
##         B  800       800
```

```
chisq.test(t)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 31.92, df = 1, p-value = 1.606e-08
```

Since the p value is very small, therefore the evidence against the null hypothesis is very strong. Hence treatment and outcome are more likely to be dependent on each other.

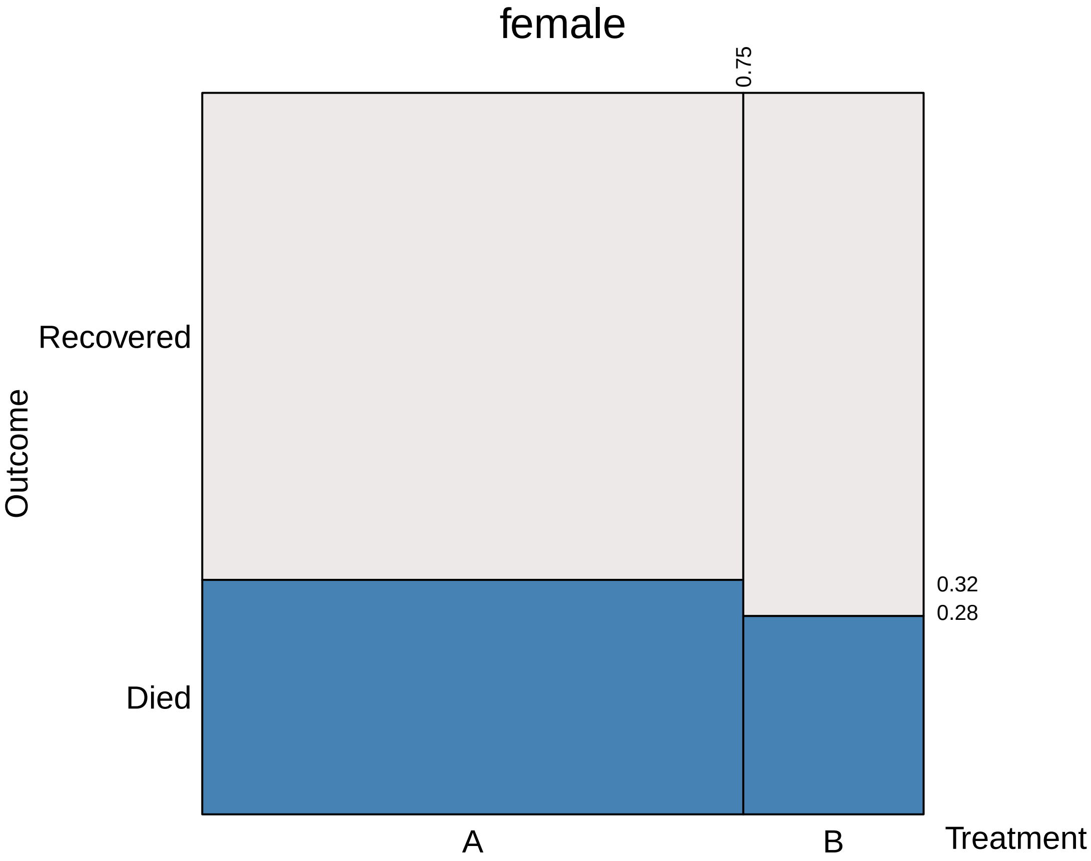
I would recommend treatment A since the number of people who recovered via A is greater than treatment

B.

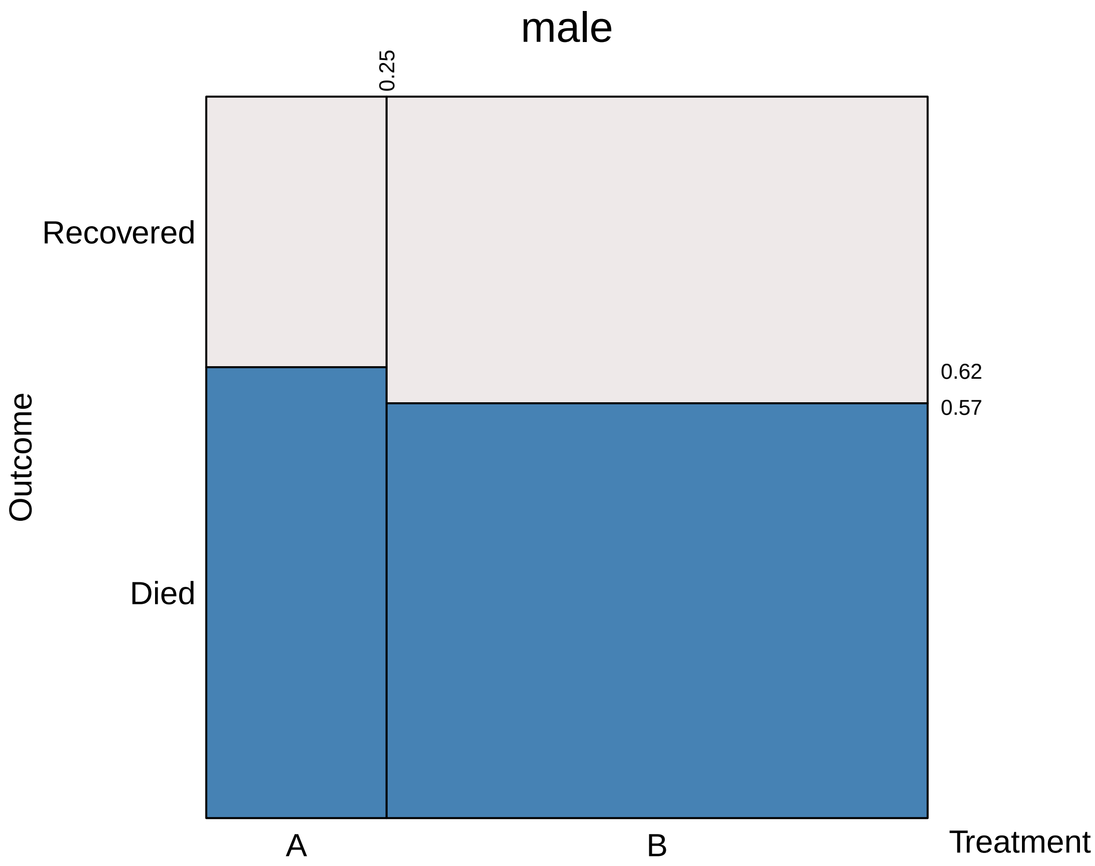
iii. *(6 marks)* There is some concern that there might be a difference between the sexes in their response to treatment. To assess this, construct the eikosograms as in part (i) above but now do so **separately for each sex**. Title each eikosogram according to the sex on which it is based.

- Show your code.
- Which sex has a greater recovery rate?
- Which treatment would this suggest for females? For males?

```
eikos(Outcome~Treatment, data = medicalRecordsTable[,1,,],main="female")
```



```
eikos(Outcome~Treatment, data = medicalRecordsTable[,2,,],main="male")
```



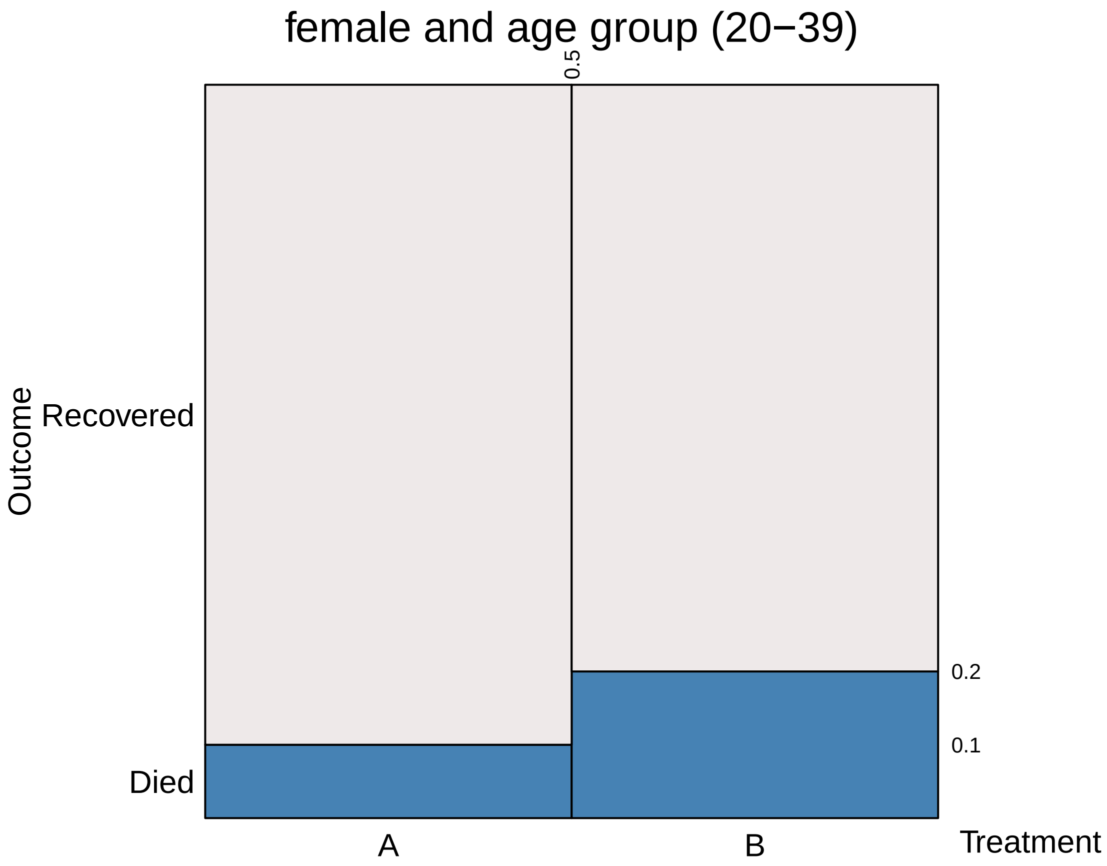
From the eikosogram, females have a greater recovery rate than males.

For both males and females, the preferred treatment would be treatment B as the outcome 'died' is less likely is when treatment is B as compared to treatment A.

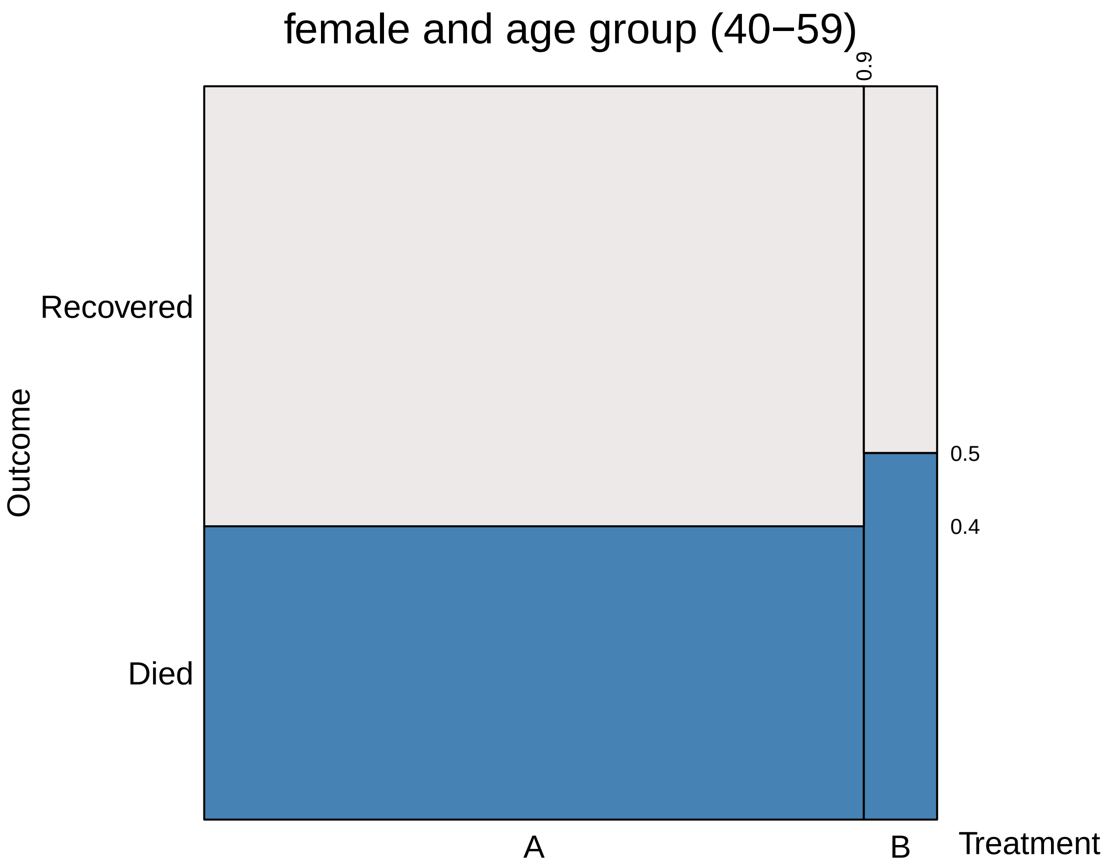
iv. *(6 marks)* Construct eikosograms as in part (iii) but now form 4 separate eikosograms, one for each combination of sex ("Male" or "Female") and age group ("20-39" and "40-59"). Title each eikosogram of by the sex and age group on which it is based.

- Show your code.
- Which treatments are now suggested for each of the four groups: Young women, older women, young men, older men?

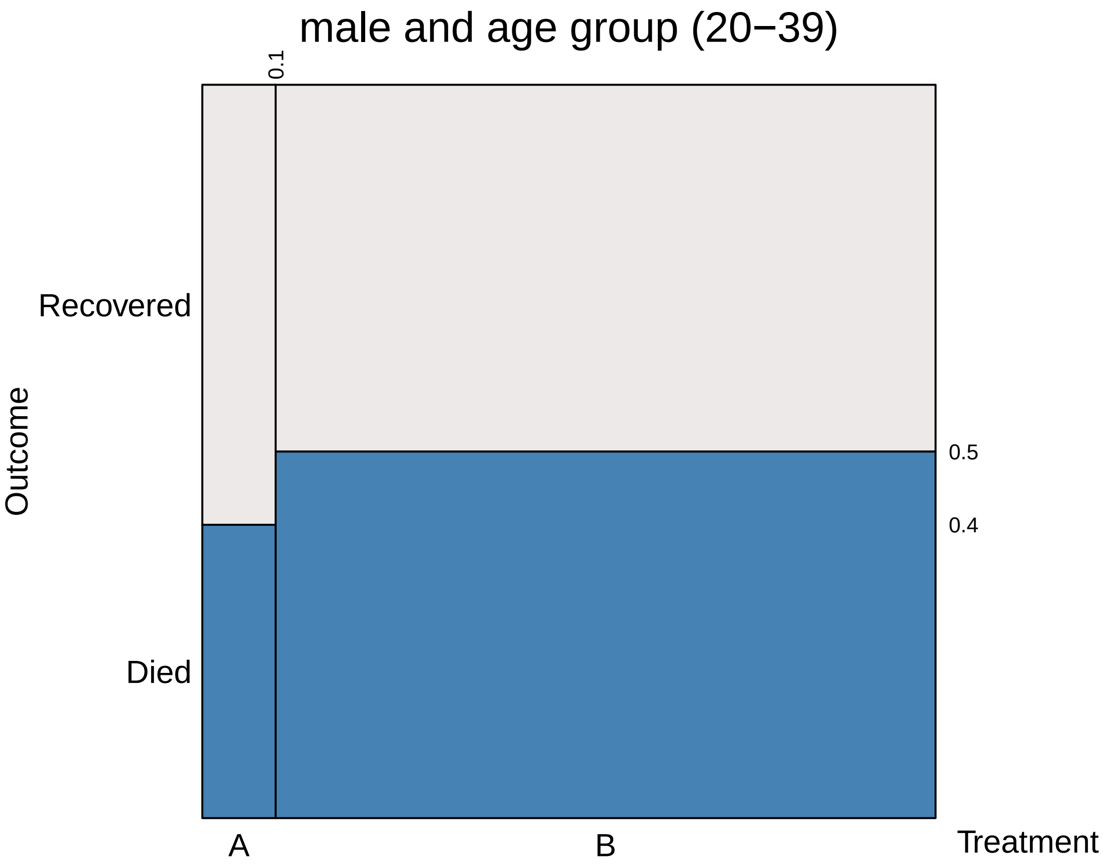
```
eikos(Outcome~Treatment, data = medicalRecordsTable[1,1,,],main="female and age group (20-39)")
```



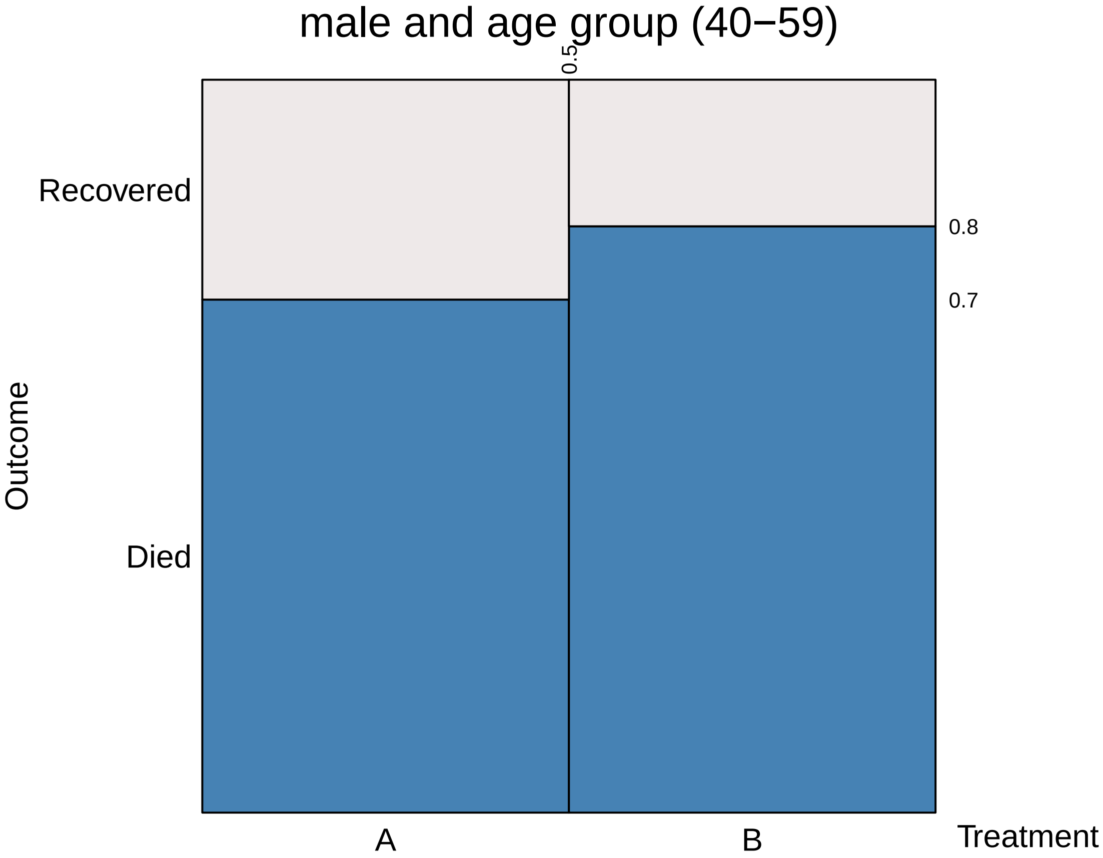
```
eikos(Outcome~Treatment, data = medicalRecordsTable[2,1,,],main="female and age group (40-59)")
```



```
eikos(Outcome~Treatment, data = medicalRecordsTable[1,2,,],main="male and age group (20-39)")
```



```
eikos(Outcome~Treatment, data = medicalRecordsTable[2,2,,],main="male and age group (40-59)")
```



For all patients irrespective of sex and age, the preferred treatment would be treatment A since the likelihood of dying is less when the given treatment is A than B.

3. The **Conclusions** stage.

i. *(4 marks)* Based on your analysis, what would you recommend to the health scientists?

Based on the analysis above, I would recommend treatment A to be given to all patients of all age groups since the outcome of death is less likely when its treatment A rather than B. Maybe data from local patients can represent the data more correctly. More patients can be treated with treatment B in the group of females (age 40-59) and more patients can be treated with treatment A in the group of males (age 20-39) to make sure which treatment is actually better since the proportion of people who got each treatment in those two groups are highly unequal.