STA 5364, Report 3.5

Carson Slater *Baylor University*

Problem

Work problem 7.1, KM p. 238.

KM 7.1 (p. 238)

In a study of the effectiveness of a combined drug regimen for the treatment of rheumatoid arthritis, 40 white patients were followed for a period ranging from 1 to 18 years. During the course of the study, 9 patients died. The ages at entry into the study and at death for these 9 patients were as follows:

Female deaths: (66, 74), (60, 76), (70, 77), (71, 81)
Male deaths: (50, 59), (60, 66), (51, 69), (69, 71), (58, 71)

For the 31 patients still alive at the end of the study their ages at entry and last follow-up were as follows:

Female Survivors: (50, 68), (55, 72), (56, 60), (45, 55), (48, 51), (44, 55), (33, 51), (44, 50), (60, 70), (55, 60), (60, 72), (77, 80), (70, 75), (66, 70), (59, 63), (62, 63)
Male Survivors: (53, 68), (55, 62), (56, 63), (45, 51), (48, 61), (49, 55), (43, 51), (44, 54), (61, 70), (45, 60), (63, 72), (74, 80), (70, 76), (66, 72), (54, 70)

Using the all-cause U.S. mortality table for 1989 (Table 2.1) test the hypothesis that the death rate of these rheumatoid arthritis patients is not different from that in the general population using the log-rank test.

Problem Description

We analyze the survival data of rheumatoid arthritis (RA) patients to test whether their mortality rate differs from that of the general population. The dataset includes:

- **9 deaths** (ages at entry and death provided).
- **31 survivors** (ages at entry and follow-up provided).

We use the U.S. all-cause mortality table for 1989 to perform the log-rank test.

Data Preparation

Step 1: Define U.S. Population Survival Data

Survival probabilities for white males and females are provided for integer ages from 0 to 85.

```
age_grid <- 0:85

# Check consistency of survival vectors
```

```r
stopifnot(length(S_WM) == length(age_grid), length(S_WF) == length(age_grid))

# Compute cumulative hazards
H0_WM <- -log(S_WM)
H0_WF <- -log(S_WF)
```

Step 2: Define Helper Functions

We create a function to compute the cumulative hazard at non-integer ages using linear interpolation.

```r
interp <- function(x, x0, x1, y0, y1) {
  y0 + ((x - x0) / (x1 - x0)) * (y1 - y0)
}

H0_func <- function(age, sex, ages, H0_male, H0_female) {
  if (age <= min(ages)) return(if (sex == "male") H0_male[1] else H0_female[1
  if (age >= max(ages)) return(if (sex == "male") H0_male[length(H0_male)] el

  a_floor <- floor(age)
  a_ceil <- ceiling(age)
  idx_floor <- which(ages == a_floor)
  idx_ceil <- which(ages == a_ceil)

  if (sex == "male") {
    interp(age, a_floor, a_ceil, H0_male[idx_floor], H0_male[idx_ceil])
  } else {
    interp(age, a_floor, a_ceil, H0_female[idx_floor], H0_female[idx_ceil])
  }
}
```

Step 3: Input RA Patient Data

We organize the RA patient data into a single data frame with survival times, status, and sex.

```r
f_deaths <- data.frame(start = c(66, 60, 70, 71), stop = c(74, 76, 77, 81), s
m_deaths <- data.frame(start = c(50, 60, 51, 69, 58), stop = c(59, 66, 69, 71

f_surv <- data.frame(
  start = c(50, 55, 56, 45, 48, 44, 33, 44, 60, 55, 60, 77, 70, 66, 59, 62),
  stop  = c(68, 72, 60, 55, 51, 55, 51, 50, 70, 60, 72, 80, 75, 70, 63, 63),
  status = 0, sex = "female"
)

m_surv <- data.frame(
  start = c(53, 55, 56, 45, 48, 49, 43, 44, 61, 45, 63, 74, 70, 66, 54),
  stop  = c(68, 62, 63, 51, 61, 55, 51, 54, 70, 60, 72, 80, 76, 72, 70),
  status = 0, sex = "male"
```

```
)

ra_data <- bind_rows(f_deaths, m_deaths, f_surv, m_surv) %>%
  mutate(sex = factor(sex, levels = c("male", "female")))
```

Log-Rank Test

Step 4: Compute Observed (O) and Expected (E) Values

Using cumulative hazard differences, calculate observed and expected counts.

```
# Female deaths: (entry, exit=death)
f_deaths_entry <- c(66, 60, 70, 71)
f_deaths_exit  <- c(74, 76, 77, 81)
f_deaths_sex   <- rep("F", 4)
f_deaths_status <- rep(1, 4)

# Male deaths: (entry, exit=death)
m_deaths_entry <- c(50, 60, 61, 69, 58)
m_deaths_exit  <- c(59, 66, 69, 71, 71)
m_deaths_sex   <- rep("M", 5)
m_deaths_status <- rep(1, 5)

# Female survivors: (entry, exit=last follow-up)
f_surv_entry <- c(50, 55, 56, 45, 48, 44, 33, 44, 60, 55, 60, 77, 70, 66, 59,
f_surv_exit  <- c(68, 72, 60, 55, 51, 55, 51, 50, 70, 60, 72, 80, 75, 70, 63,
f_surv_sex   <- rep("F", 16)
f_surv_status <- rep(0, 16)

# Male survivors: (entry, exit=last follow-up)
m_surv_entry <- c(53, 55, 56, 45, 48, 49, 43, 44, 61, 45, 63, 74, 70, 66, 54)
m_surv_exit  <- c(68, 62, 63, 51, 61, 55, 51, 54, 70, 60, 72, 80, 76, 72, 74)
m_surv_sex   <- rep("M", 15)
m_surv_status <- rep(0, 15)

# Combine all data
entry_age <- c(f_deaths_entry, m_deaths_entry,
               f_surv_entry, m_surv_entry)
exit_age <- c(f_deaths_exit, m_deaths_exit,
              f_surv_exit, m_surv_exit)
sex <- c(f_deaths_sex, m_deaths_sex,
         f_surv_sex, m_surv_sex)
status <- c(f_deaths_status, m_deaths_status,
            f_surv_status, m_surv_status)

# Create an ID vector
```

```r
id <- 1:length(entry_age)

# Construct the data frame
ra_data <- data.frame(
  id = id,
  sex = sex,
  entry_age = entry_age,
  exit_age = exit_age,
  status = status
)



# Define a helper function to get population survival by sex
get_pop_surv <- function(age, sex) {
  age_int <- floor(age)
  if (sex == "M") {
    return(S_WM[pmin(age_int+1, length(S_WM))]) # handle edge if needed
  } else {
    return(S_WF[pmin(age_int+1, length(S_WF))])
  }
}

# Compute expected deaths per patient
ra_data$expected_death <- 0
for (i in 1:nrow(ra_data)) {
  ent <- floor(ra_data$entry_age[i])
  ex <- floor(ra_data$exit_age[i])
  s <- ra_data$sex[i]

  # sum q(a) = S(a) - S(a+1) over each year interval patient was at risk
  # approximate since we only have integer ages:
  E <- 0
  for (a in ent:(ex-1)) {
    # Probability of surviving to age a and a+1:
    if (s == "M") {
      Sa <- S_WM[a+1]
      Sa1 <- S_WM[a+2]
    } else {
      Sa <- S_WF[a+1]
      Sa1 <- S_WF[a+2]
    }
    q <- Sa - Sa1
    E <- E + q
  }
  ra_data$expected_death[i] <- E
}
```

```r
# Total observed and expected
O <- sum(ra_data$status)
E <- sum(ra_data$expected_death)


ra_surv <- Surv(time = ra_data$exit_age - ra_data$entry_age, event = ra_data$


ra_data$group <- "RA"

pseudo_data <- ra_data
pseudo_data$status <- 0
pseudo_data$group <- "Expected"

combined <- rbind(
  data.frame(time=ra_data$exit_age - ra_data$entry_age, status=ra_data$status
  data.frame(time=pseudo_data$exit_age - pseudo_data$entry_age, status=pseudo_
)
```

Step 5: Calculate Test Statistic and p-Value

```r
# Now run survdiff:
res <- survdiff(Surv(time, status) ~ group, data=combined)
res
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ group, data = combined)
##
##                  N Observed Expected
## group=Expected  40        0      4.5
## group=RA        40        9      4.5
##                 (O-E)^2/E (O-E)^2/V
## group=Expected        4.5      9.05
## group=RA              4.5      9.05
##
##  Chisq= 9  on 1 degrees of freedom, p= 0.003
```

```r
cat("Test statistic (U_LR):", res$chisq, "\n")
```

```
## Test statistic (U_LR): 9.049046
```

```r
cat("p-value:", round(res$pvalue, 3), "\n")
```

```
## p-value: 0.003
```

```r
if (round(res$pvalue, 3) < 0.05) {
  cat("Conclusion: Reject the null hypothesis. RA patients have a different mo
} else {
  cat("Conclusion: Fail to reject the null hypothesis. No evidence of a diffe
}
```

```
## Conclusion: Reject the null hypothesis. RA patients have a different mortal
```

Results and Interpretation

The log-rank test allows us to compare the observed and expected deaths to determine if the RA patients' mortality rate significantly differs from that of the general population. A small p-value (< 0.05) indicates a significant difference.