LeaRning about Statistics

Tables, Chi Square and Fisher's exact test, KM curves and log rank tests

Zachary Thompson
Moffitt Cancer Center

July 12, 2022

What you will learn to run

- Review of tables with janitor package
- Chi Square and Fisher's Exact test
- Kaplan-Meier Curves and estimates
- Log rank tests

Data prep: location use here()

```
here()
```

```
[1] "F:/myGitRepo/Intro_to_R_2022"
```

```
here("data","tcga-clinical.txt")
```

```
[1] "F:/myGitRepo/Intro_to_R_2022/data/tcga-clinical.txt"
```

```
here("data","tcga-gene-exp.txt")
```

```
[1] "F:/myGitRepo/Intro_to_R_2022/data/tcga-gene-exp.txt"
```

Data prep: load

```
clinical <- read.csv(file = here("data","tcga_clinical.txt"),
                     header = TRUE)
geneexp <- read.csv(file = here("data","tcga_gene_exp.txt"),
                    header = TRUE)
```

Data prep: merge

```
tcga <- left_join(clinical, geneexp, by = "bcr_patient_barcode")
```

```
intersect(names(clinical), names(geneexp))
```

```
[1] "bcr_patient_barcode"
```

```
dim(tcga)
```

```
[1] 1043   33
```

Data prep: mutate

```
  tcga <- tcga %>% mutate(
  smoking = case_when(
    tobacco_smoking_history %in% c(
      "Current reformed smoker for < or = 15 years",
      "Current reformed smoker for > 15 years",
      "Current Reformed Smoker, Duration Not Specified"
    ) ~ "Former",
    tobacco_smoking_history %in% c("Current smoker") ~ "Current",
    tobacco_smoking_history %in% c("Lifelong Non-smoker") ~ "Never",
    is.na(tobacco_smoking_history) ~ NA_character_
  )
)
```

Review of tables

```
janitor::tabyl(tcga, radiation_therapy, vital_status )
```

```
radiation_therapy Alive Dead
               NO   217   84
              YES   190   95
             <NA>   243  214
```

Review of tables

```
janitor::tabyl(tcga, radiation_therapy, vital_status,
               show_na = FALSE )
```

```
radiation_therapy Alive Dead
               NO   217   84
              YES   190   95
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE )
```

```
smoking FEMALE MALE
Current     40  151
 Former     54  180
  Never     72   88
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row"))
```

```
smoking FEMALE MALE
Current     40  151
 Former     54  180
  Never     72   88
  Total    166  419
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("col"))
```

```
smoking FEMALE MALE Total
Current     40  151   191
 Former     54  180   234
  Never     72   88   160
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col"))
```

```
smoking FEMALE MALE Total
Current     40  151   191
 Former     54  180   234
  Never     72   88   160
  Total    166  419   585
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "col")
```

```
smoking    FEMALE      MALE     Total
Current 0.2409639 0.3603819 0.3264957
 Former 0.3253012 0.4295943 0.4000000
  Never 0.4337349 0.2100239 0.2735043
  Total 1.0000000 1.0000000 1.0000000
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "col") %>%
  adorn_pct_formatting(digits = 0)
```

```
smoking FEMALE MALE Total
Current    24%  36%   33%
 Former    33%  43%   40%
  Never    43%  21%   27%
  Total   100% 100%  100%
```

Review of tables

```
tcga %>%
  tabyl(smoking, gender, show_na = FALSE ) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "col") %>%
  adorn_pct_formatting(digits = 0) %>%
  adorn_ns(position = "front")
```

smoking	FEMALE		MALE		Total	
Current	40	(24%)	151	(36%)	191	(33%)
Former	54	(33%)	180	(43%)	234	(40%)
Never	72	(43%)	88	(21%)	160	(27%)
Total	166	(100%)	419	(100%)	585	(100%)

Pearson's chi-squared test

The chi squared test is a non-parametric test that can be applied to contingency tables with various dimensions. The name of the test originates from the chi-squared distribution, which is the distribution for the squares of independent standard normal variables. This is the distribution of the test statistic of the chi squared test, which is defined by the sum of chi-square values for all cells arising from the difference between a cell's observed value and the expected value, normalized by the expected value.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- χ^2 = chi square statistic
- O_{ij} = observed value
- E_{ij} = expected value

Pearson’s chi-squared test

The null hypothesis of the Chi-Square test is that no relationship exists between the categorical variables in the population; they are independent.

smoking	FEMALE		MALE		Total	
Current	40	(24%)	151	(36%)	191	(33%)
Former	54	(33%)	180	(43%)	234	(40%)
Never	72	(43%)	88	(21%)	160	(27%)
Total	166	(100%)	419	(100%)	585	(100%)

Pearson’s chi-squared test

The null hypothesis of the Chi-Square test is that no relationship exists between the categorical variables in the population; they are independent.

```
tcga %>% tabyl( smoking, gender , show_na = FALSE ) %>%
  chisq.test()
```

```
    Pearson's Chi-squared test

data:  .
X-squared = 30.182, df = 2, p-value = 0.0000002793
```

The p-value is very low so we reject the null hypothesis that there is no association between the variables.

Pearson's chi-squared test

More women are never smokers and more men are current smokers.

R Code | Plot

```
ggplot(tcga %>% filter(!is.na(smoking)),
       aes(x = gender, fill = smoking )) +
  geom_bar(position = "fill") +
  labs(y = "proportion")
```

Pearson’s chi-squared test

More women are never smokers and more men are current smokers.

R Code | Plot

Fisher’s Exact test

Similar to Chi square test. Use the Fisher's exact test of independence when you have two nominal variables and you want to see whether the proportions of one variable are different depending on the value of the other variable. Use it when the sample size is small.

How the test works:

Unlike most statistical tests, Fisher's exact test does not use a mathematical function that estimates the probability of a value of a test statistic; instead, you calculate the probability of getting the observed data, and all data sets with more extreme deviations, under the null hypothesis that the proportions are the same.

Fisher's Exact test

```
tcga <- tcga  %>% mutate(Death = ifelse(vital_status == "Dead"," Yes","I

tcga %>% filter(race %in% c("ASIAN")) %>%
  tabyl(gender, Death, show_na = TRUE) %>%
  adorn_totals(where = c("row","col")) %>%
  adorn_percentages(denominator = "row") %>%
  adorn_pct_formatting(digits = 0) %>%
  adorn_ns(position = "front")
```

gender	Yes	No	Total
FEMALE	2 (25%)	6 (75%)	8 (100%)
MALE	3 (27%)	8 (73%)	11 (100%)
Total	5 (26%)	14 (74%)	19 (100%)

Fisher's Exact test

The null hypothesis is that the relative proportions of one variable are independent of the second variable. For 2 by 2 tables, the null of conditional independence is equivalent to the hypothesis that the odds ratio equals one.

```
tcga %>% filter(race %in% c("ASIAN")) %>%
  tabyl(gender, Death, show_na = TRUE) %>% fisher.test()
```

```
    Fisher's Exact Test for Count Data

data:  .
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.05723439 10.70637705
sample estimates:
odds ratio
 0.8943933
```

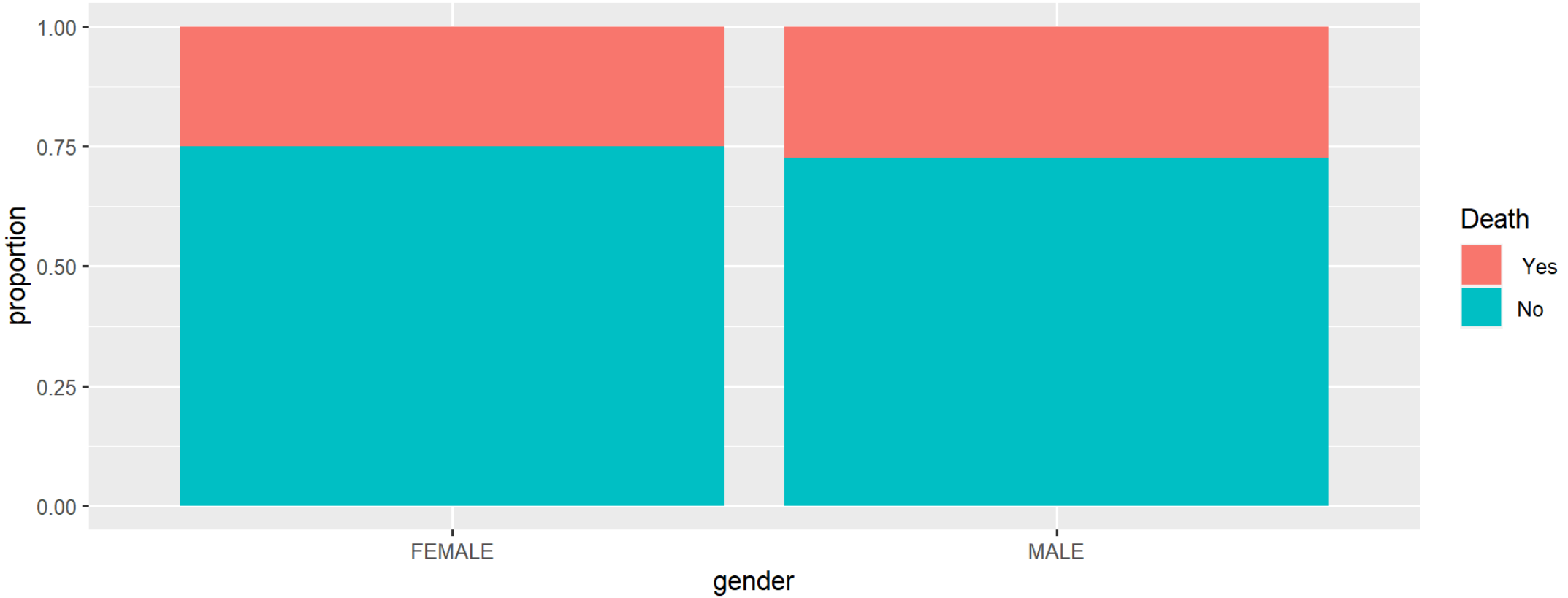
Fisher's Exact test

R Code | Plot

```
ggplot(tcga %>% filter(race %in% c("ASIAN" )),
       aes(x = gender, fill = Death )) +
  geom_bar(position = "fill") +
  labs(y = "proportion")
```

Fisher's Exact test

R Code | Plot



Time to event data - survival data

Time-to-event data consist of pairs of observations for each individual:

- (i) a length of time during which no event was observed
- (ii) an indicator of whether the end of that time period corresponds to an event or just the end of observation
- Survival ~ event is death (OS)
- Progression free survival ~ event is death or disease progression (PFS)

Kaplan-Meier estimates and curves

The Kaplan–Meier estimator, also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from time to event data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. In other fields, Kaplan–Meier estimators may be used to measure the time-to-failure of machine parts.

Kaplan-Meier estimates

The estimator of the survival function:

$$\hat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right)$$

- $\hat{S}(t)$ = probability survival is longer than t
- d_i = number of deaths (events) that happened at time t_i
- n_i = number of individuals that have survived to time t_i

Kaplan-Meier estimates and curves

A plot of the Kaplan–Meier estimator is a series of declining horizontal steps which, with a large enough sample size, approaches the true survival function for that population.

An important advantage of the Kaplan–Meier curve is that the method can take into account some types of censored data, particularly right-censoring, which occurs if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up. On the plot, small vertical tick-marks state individual patients whose survival times have been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is the complement of the empirical distribution function.

Kaplan-Meier estimates and curves: Data

R Code | R Result

```
OSdata <-  data.frame(OStime = tcga$OS.time/365,
                      Death = I(tcga$vital_status=="Dead"))
```

Kaplan-Meier estimates and curves: Data

R Code | R Result

```
head(OSdata, 10)
```

```
       OStime Death
1  1.05479452 FALSE
2  0.27945205 FALSE
3  0.99178082 FALSE
4  3.06849315 FALSE
5  3.93424658 FALSE
6  0.04383562 FALSE
7  3.26301370  TRUE
8  2.01369863  TRUE
9  4.09041096 FALSE
10 4.08493151 FALSE
```

Kaplan-Meier estimates and curves: Survival object

R Code | R Result

```
OS_obj<-survival::Surv(tcga$OS.time/365, I(tcga$vital_status=="Dead"))
```

Kaplan-Meier estimates and curves: Survival object

R Code | R Result

```
head(OS_obj,40)
```

```
 [1] 1.05479452+ 0.27945205+ 0.99178082+ 3.06849315+ 3.93424658+
 [6] 0.04383562+ 3.26301370  2.01369863  4.09041096+ 4.08493151+
[11] 3.09589041+ 4.13150685+ 4.04931507+ 3.03013699+ 3.24931507+
[16] 3.20547945  3.11506849+ 3.79452055+ 4.44931507+ 4.06849315+
[21] 5.16712329+ 2.58904110+ 0.37534247  4.41095890  3.79452055+
[26] 2.49315068+ 1.53698630  7.36438356+ 3.58082192+ 6.86027397+
[31] 4.27123288+ 0.87397260+ 2.39178082+ 6.21917808+ 6.23013699+
[36] 2.01369863+ 4.44109589+ 3.60000000+ 4.64657534  1.72602740+
```

Kaplan-Meier estimates and curves: Data again

R Code | R Result

```
OSdata <- bind_cols(OSdata, OS_obj)
```

Kaplan-Meier estimates and curves: Data again

R Code | R Result

```
head(OSdata, 9)
```

```
      OStime Death       time status
1 1.05479452 FALSE 1.05479452      0
2 0.27945205 FALSE 0.27945205      0
3 0.99178082 FALSE 0.99178082      0
4 3.06849315 FALSE 3.06849315      0
5 3.93424658 FALSE 3.93424658      0
6 0.04383562 FALSE 0.04383562      0
7 3.26301370  TRUE 3.26301370      1
8 2.01369863  TRUE 2.01369863      1
9 4.09041096 FALSE 4.09041096      0
```

Kaplan-Meier estimates and curves: Survfit

R Code | R Result

```
kmfit <-  survfit(OS_obj ~ gender, data = tcga)
```

Kaplan-Meier estimates and curves: Survfit

R Code | R Result

```
kmfit
```

```
Call: survfit(formula = OS_obj ~ gender, data = tcga)

   1 observation deleted due to missingness
                n events median 0.95LCL 0.95UCL
gender=FEMALE 323    133   5.93    4.54    7.54
gender=MALE   719    260   6.18    5.44    7.75
```

Kaplan-Meier estimates and curves: Survfit

```
summary(kmfit)
```

```
Call: survfit(formula = OS_obj ~ gender, data = tcga)

1 observation deleted due to missingness
                gender=FEMALE
     time n.risk n.event survival std.err lower 95% CI upper 95% CI
  0.00548    321       1   0.9969 0.00311        0.9908        1.000
  0.06301    313       1   0.9937 0.00444        0.9850        1.000
  0.08767    311       1   0.9905 0.00546        0.9799        1.000
  0.13973    310       1   0.9873 0.00631        0.9750        1.000
  0.15342    309       1   0.9841 0.00705        0.9704        0.998
  0.17534    307       1   0.9809 0.00772        0.9659        0.996
  0.17808    306       1   0.9777 0.00833        0.9615        0.994
  0.18630    305       1   0.9745 0.00890        0.9572        0.992
  0.20000    304       1   0.9713 0.00943        0.9530        0.990
  0.21096    303       1   0.9681 0.00993        0.9488        0.988
  0.24658    302       1   0.9649 0.01040        0.9447        0.985
```

Kaplan-Meier estimates and curves: Estimating x-year survival

```
summary(survfit(OS_obj ~ gender, data = tcga), times = 5, extend=TRUE)
```

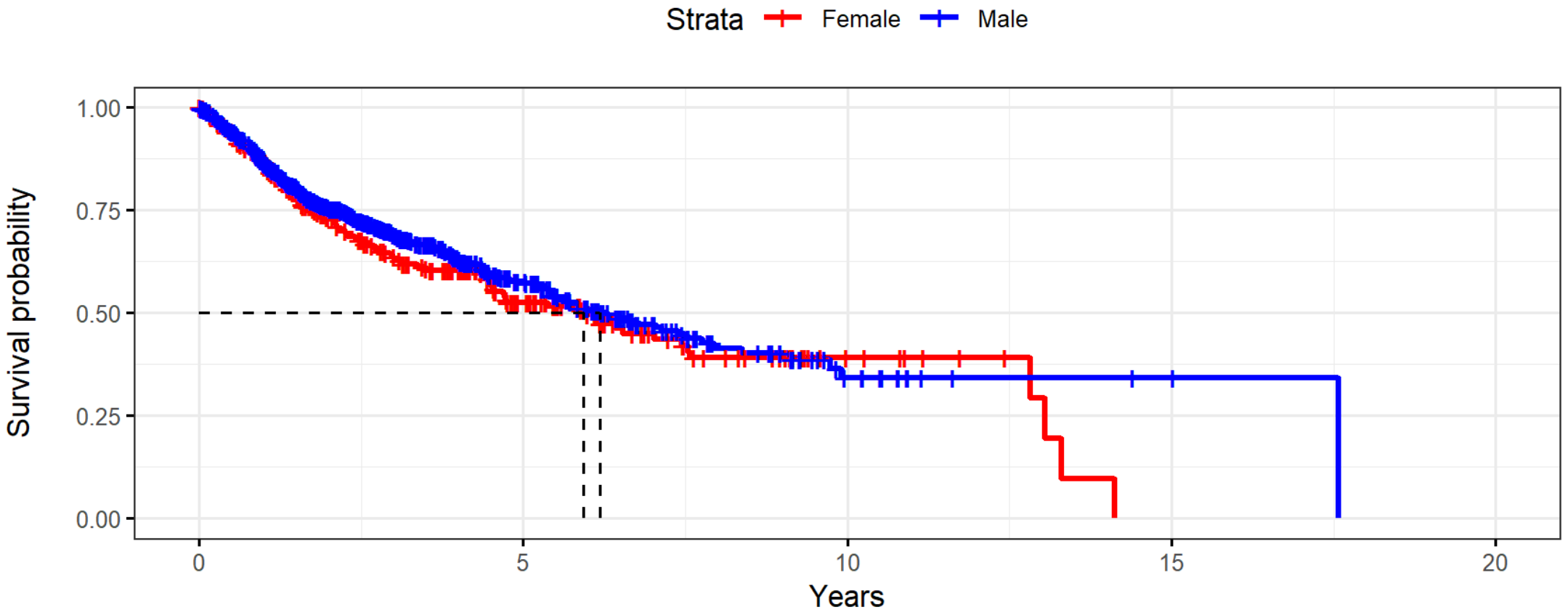
```
Call: survfit(formula = OS_obj ~ gender, data = tcga)

1 observation deleted due to missingness
                gender=FEMALE
        time         n.risk      n.event     survival      std.err
      5.0000      70.0000     118.0000       0.5249       0.0341
lower 95% CI upper 95% CI
      0.4622       0.5961

                gender=MALE
        time         n.risk      n.event     survival      std.err
      5.0000     135.0000     230.0000       0.5759       0.0236
lower 95% CI upper 95% CI
      0.5314       0.6240
```

Plot code

p <- ggsurvplot(kmfit, data = tcga,

```
        # Add median survival lines
        surv.median.line = "hv",
        # Change legends: title & labels
        title= "Overall Survival by gender",
        xlab ="Years", legend.labs = c("Female","Male"  ),
        # Add p-value and CIs
        pval = FALSE,
        conf.int = FALSE,
        # Add risk table
        risk.table = TRUE, tables.height = 0.2,
        tables.theme = theme_cleantable(),
        # Color palettes.
        palette = c("Red","Blue" )
        # Change ggplot2 theme
        ggtheme = theme_bw(),
        risk.table.title = "Number at risk")
```

Overall Survival by gender



Number at risk

Female	323	70	10	0	0
Male	719	135	14	2	0

Log rank test

The log rank test is a hypothesis test to compare the survival distributions of two or more samples.

- Nonparametric test and appropriate to use when the data are right skewed and censored.
- Widely used in clinical trials to establish the efficacy of a new treatment in comparison with a control treatment when the measurement is the time to event.

The test is sometimes called the Mantel–Cox test, named after Nathan Mantel and David Cox. The logrank test can also be viewed as a time-stratified Cochran–Mantel–Haenszel test.

Log rank test

The logrank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event (death) at any time point. The analysis is based on the times of events (deaths). For each such time we calculate the observed number of deaths in each group and the number expected if there were in reality no difference between the groups.

Log rank test: test statistic

We can now use a χ^2 test of the null hypothesis.

Test statisic = $\frac{\sum_{it}(O_{ij}-E_{ij})^2}{var(\sum_1^k(O_{ij}-E_{ij}))}$ for each group, where O and E are the totals of the observed and expected events at the time of each event.

- i, denotes the group; j denotes the time that the event occurred,
- O_{ij}, number of observed events in the ith group at the jth time period,
- E_{ij}, number of expected events in the ith group at the j th time period.

Log rank test: test statistic

- $O_{ij} = \sum_{j=1}^{K} d_{ij}$
- $E_{ij} = \sum_{j=1}^{K} d_{ij} \frac{r_{ij}}{r_j}$

Test statistic = $\frac{(\sum_j (d_{ij} - d_j \frac{r_{ij}}{r_j}))^2}{\sum_{j=1}^{k} \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}}$

Log rank test

The logrank test is based on the same assumptions as the Kaplan Meier survival curve, namely, that censoring is unrelated to prognosis, the survival probabilities are the same for subjects recruited early and late in the study, and the events happened at the times specified. Deviations from these assumptions matter most if they are satisfied differently in the groups being compared, for example if censoring is more likely in one group than another.

Because the logrank test is purely a test of significance it cannot provide an estimate of the size of the difference between the groups or a confidence interval.

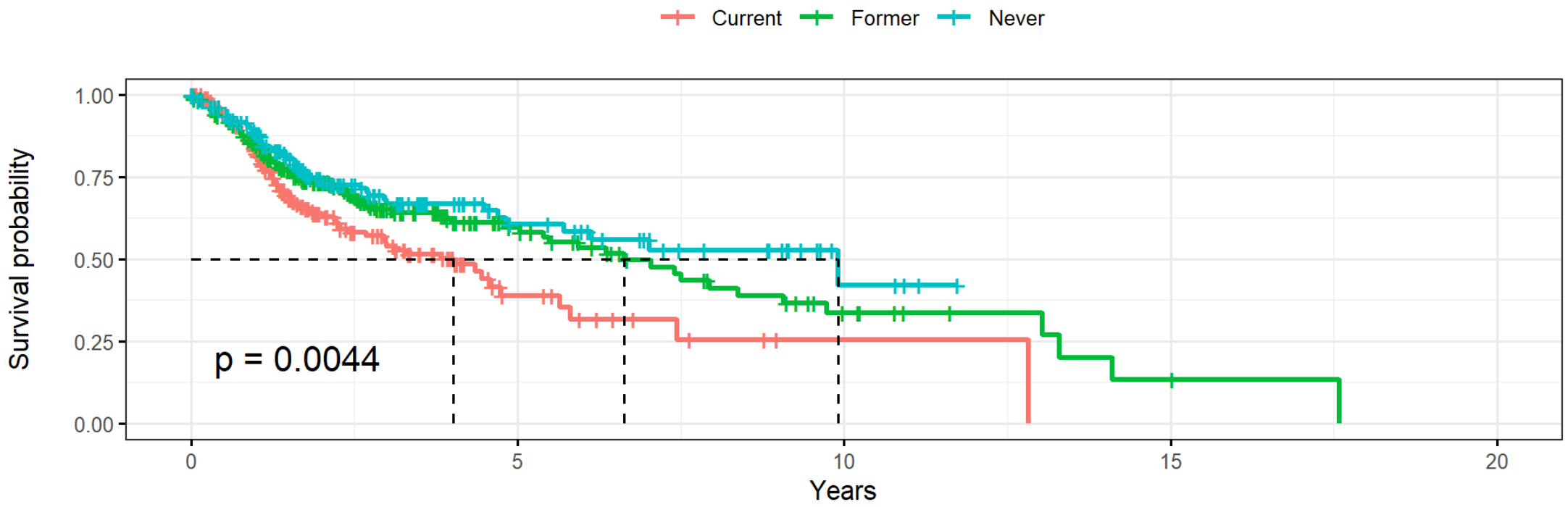
Log rank test R code

```
survdiff(OS_obj ~ smoking, data = tcga)
```

```
Call:
survdiff(formula = OS_obj ~ smoking, data = tcga)

n=584, 459 observations deleted due to missingness.

                   N Observed Expected (O-E)^2/E (O-E)^2/V
smoking=Current 190       82     61.3     7.017     9.945
smoking=Former  234       92     98.6     0.444     0.823
smoking=Never   160       49     63.1     3.157     4.463

 Chisq= 10.8  on 2 degrees of freedom, p= 0.004
```

Kaplan-Meier plots

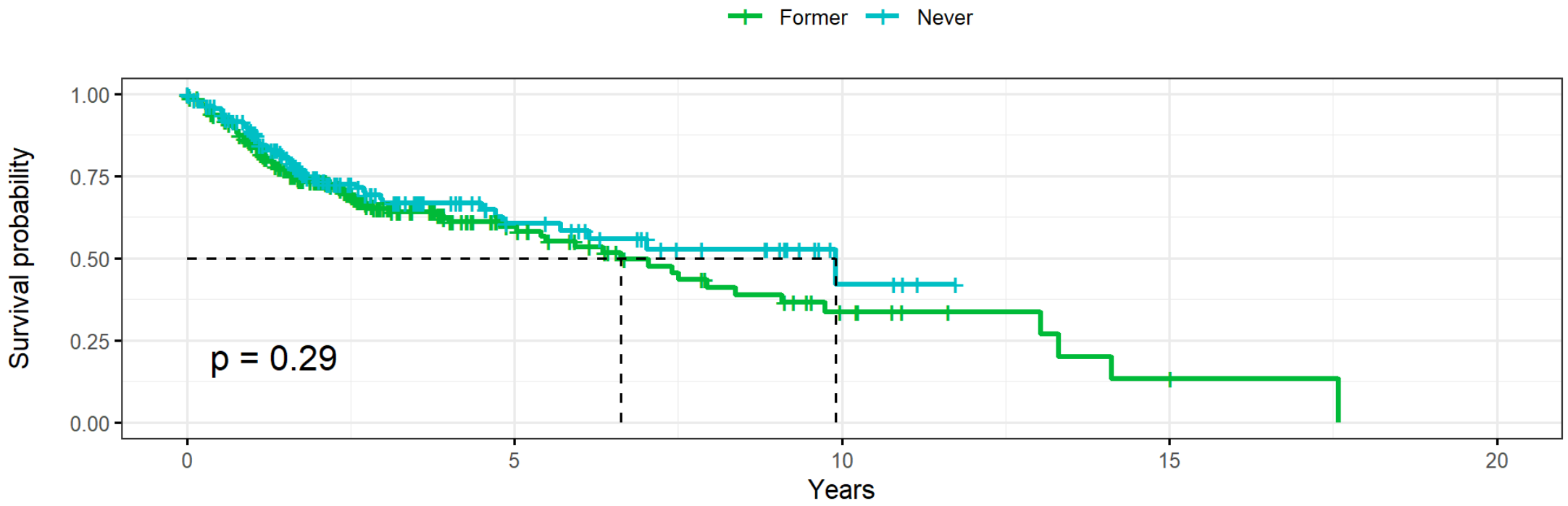
Overall Survival by smoking status



Number at risk

Current	190	13	1	0	0
Former	234	42	10	2	0
Never	160	28	4	0	0

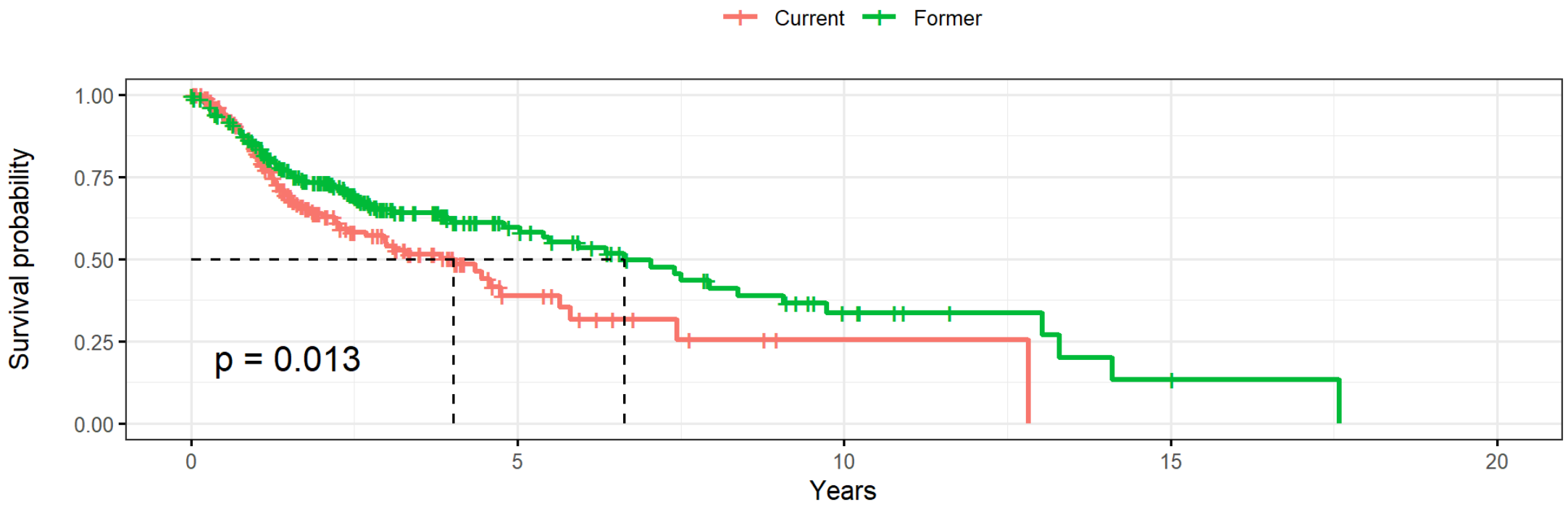
Overall Survival by smoking status - Former vs Never



Number at risk

Former	234	42	10	2	0
Never	160	28	4	0	0

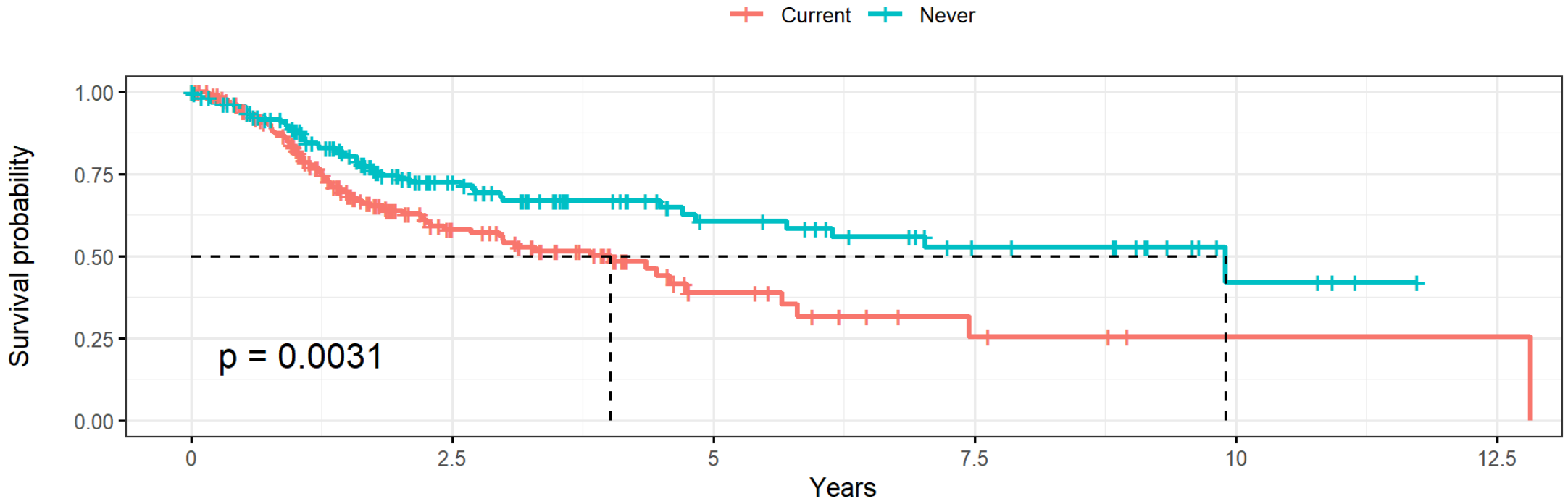
Overall Survival by smoking status - Current vs Former



Number at risk

Current	190	13	1	0	0
Former	234	42	10	2	0

Overall Survival by smoking status - Current vs Never



Number at risk

Current	190	56	13	4	1	1
Never	160	66	28	15	4	0

Log rank test R code: pair wise comparisions

R Code | R Result

```
library(survminer)

tcganonasmoking <- tcga %>% filter(!is.na(smoking)) %>%
  mutate(  OS_objs = survival::Surv( OS.time/365, I( vital_status=="Dea
```

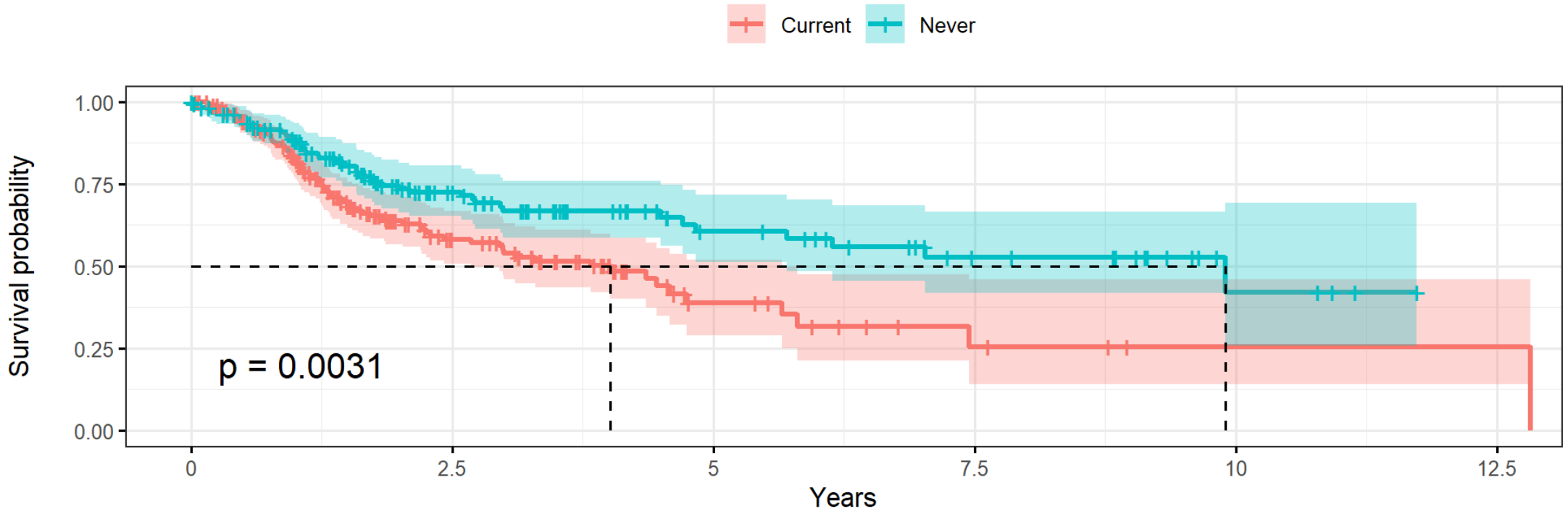
Log rank test R code: pair wise comparisions

R Code | R Result

```
pairwise_survdiff(OS_objs ~  smoking, data = tcganonasmoking,
                  p.adjust.method = "none", rho = 0)
```

```
    Pairwise comparisons using Log-Rank test

data:  tcganonasmoking and smoking

       Current Former
Former 0.0131  -
Never  0.0031  0.2900

P value adjustment method: none
```

Overall Survival by smoking status - Current vs Never



Number at risk

Current	190	56	13	4	1	1
Never	160	66	28	15	4	0

Thank you!

- The end