Life Table estimate:Simulation and Real Dataset Study

Parna Dutta,Sampriti Dey, Sukanya Pal

2025-12-09

Introduction

The *Life Table* method estimates survival by dividing follow-up time into intervals and computing survival probabilities using the number at risk, events, and censored observations. Because censoring times are not exact, it relies on assumptions—typically that censored individuals withdraw uniformly within each interval. These assumptions and the grouping of event times introduce approximation error and potential bias, especially when intervals are wide or censoring is uneven. In contrast, the *Kaplan–Meier estimator* uses exact event and censoring times, avoids interval grouping, and therefore provides a more accurate and unbiased survival estimate. For this reason, Kaplan–Meier is preferred whenever precise event times are available, while the Life Table method is mainly used for grouped or older datasets.

Methodology

The study will begin by simulating time-to-event data where both event times and cen- soring times are generated from EXPONENTIAL DISTRIBUTION under controlled settings. Withdrawal patterns will be artificially imposed—assuming uniform distri- bution of withdrawals across intervals, concentration of withdrawals at the beginning, and concentration at the end of intervals. Life Table survival estimates will be com- puted using each of the three formula variations. These estimates will be compared against the “true” simulated survival function to evaluate bias and variability.

In the empirical part of the project, the *kidney* dataset from the “KMsurv” package will be used. This dataset contains 119 observations on kidney infection recurrence times, event indicators, and treatment types. Survival times will be grouped into intervals, and life tables will be constructed. Kaplan–Meier estimates will also be computed for comparison. Graphical and numerical summaries will be used to study the differences between the methods.

Loading Packages

```
library(tidyverse)
library(survival)
library(KMsurv)
library(dplyr)
```

1. Simulation Study

1.1 Life-Table Builder Function

```
build_life_table <- function(df, breaks) {
df <- df %>% mutate(interval = cut(time, breaks = breaks, right = TRUE, include.lowest = TRUE))
lt <- df %>%
group_by(interval) %>%
summarise(
d_j = sum(status == 1 , na.rm = TRUE),
w_j = sum(status == 0 , na.rm = TRUE),
inside = n()
) %>% ungroup()

n_j <- numeric(nrow(lt))
n_j[1] <- nrow(df)
if(nrow(lt) >= 2) {
for(j in 2:nrow(lt)) {
n_j[j] <- n_j[j-1] - lt$d_j[j-1] - lt$w_j[j-1]
}
}

lt <- lt %>% mutate(
n_j = n_j,
denom_uniform = n_j - w_j/2,
q_uniform = ifelse(denom_uniform > 0, d_j / denom_uniform, NA),
q_early   = ifelse((n_j - w_j) > 0, d_j / (n_j - w_j), NA),
q_late    = ifelse(n_j > 0, d_j / n_j, NA)
)

lt <- lt %>% mutate(
S_uniform = cumprod(1 - replace_na(q_uniform, 0)),
S_early   = cumprod(1 - replace_na(q_early, 0)),
S_late    = cumprod(1 - replace_na(q_late, 0))
)

interval_ends <- as.numeric(gsub("\\((.+),(.+)\\]", "\\2", lt$interval))
lt <- lt %>% mutate(time = interval_ends)
return(lt)
}
```

1.2 Simulation of Survival Data

```
set.seed(2025)
n <- 1000
lambdaT <- 0.1
lambdaC <- 0.05

T <- rexp(n, rate = lambdaT)
C <- rexp(n, rate = lambdaC)

time <- pmin(T, C)
status <- as.integer(T <= C)

s_df <- data.frame(time, status)
```

```
head(s_df)
```

```
##          time status
## 1  4.65240310      1
## 2 10.38634094      1
## 3  5.60569047      1
## 4  0.08504318      1
## 5  0.88620726      1
## 6  1.39062109      1
```

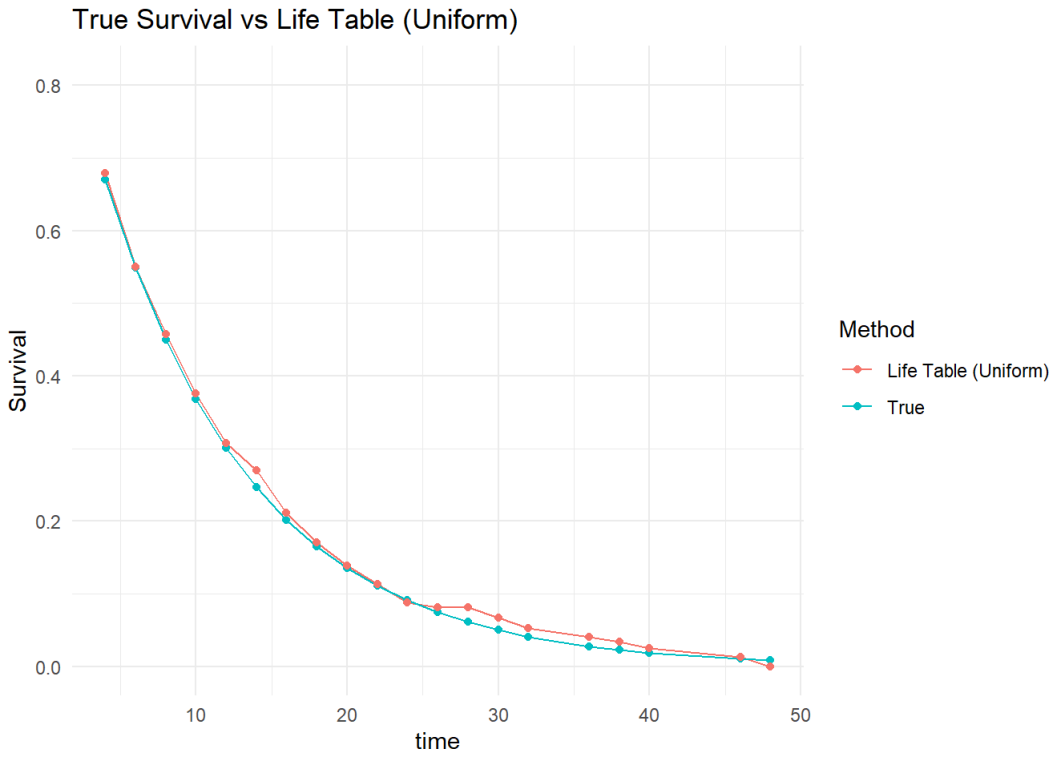
1.3 Life Table (Simulated Data)

```
breaks <- seq(0, max(s_df$time) + 2, by = 2)
lt_sim <- build_life_table(s_df, breaks)
lt_sim
```

```
## # A tibble: 21 × 13
##    interval   d_j   w_j inside   n_j denom_uniform q_uniform q_early q_late
##    <fct>    <int> <int>  <int> <dbl>         <dbl>     <dbl>   <dbl>  <dbl>
##  1 [0,2]      178    84    262  1000          958       0.186   0.194  0.178
##  2 (2,4]      117    60    177   738          708       0.165   0.173  0.159
##  3 (4,6]      103    41    144   561          540.      0.191   0.198  0.184
##  4 (6,8]       67    36    103   417          399       0.168   0.176  0.161
##  5 (8,10]      54    29     83   314          300.      0.180   0.189  0.172
##  6 (10,12]     40    20     60   231          221       0.181   0.190  0.173
##  7 (12,14]     20    13     33   171          164.      0.122   0.127  0.117
##  8 (14,16]     29    11     40   138          132.      0.219   0.228  0.210
##  9 (16,18]     18     8     26    98           94       0.191   0.2    0.184
## 10 (18,20]     13     7     20    72           68.5     0.190   0.2    0.181
## # ℹ 11 more rows
## # ℹ 4 more variables: S_uniform <dbl>, S_early <dbl>, S_late <dbl>, time <dbl>
```

1.4 True Survival vs Life Table Plot

```
interval_ends <- lt_sim$time
S_true <- exp(-(lambdaT * interval_ends))

plot_df <- data.frame(
time = rep(interval_ends, 2),
Survival = c(S_true, lt_sim$S_uniform),
Method = rep(c("True", "Life Table (Uniform)"), each = length(interval_ends))
)

ggplot(plot_df, aes(x = time, y = Survival, color = Method)) +
geom_line() +
geom_point() +
theme_minimal() +
labs(title = "True Survival vs Life Table (Uniform)")
```

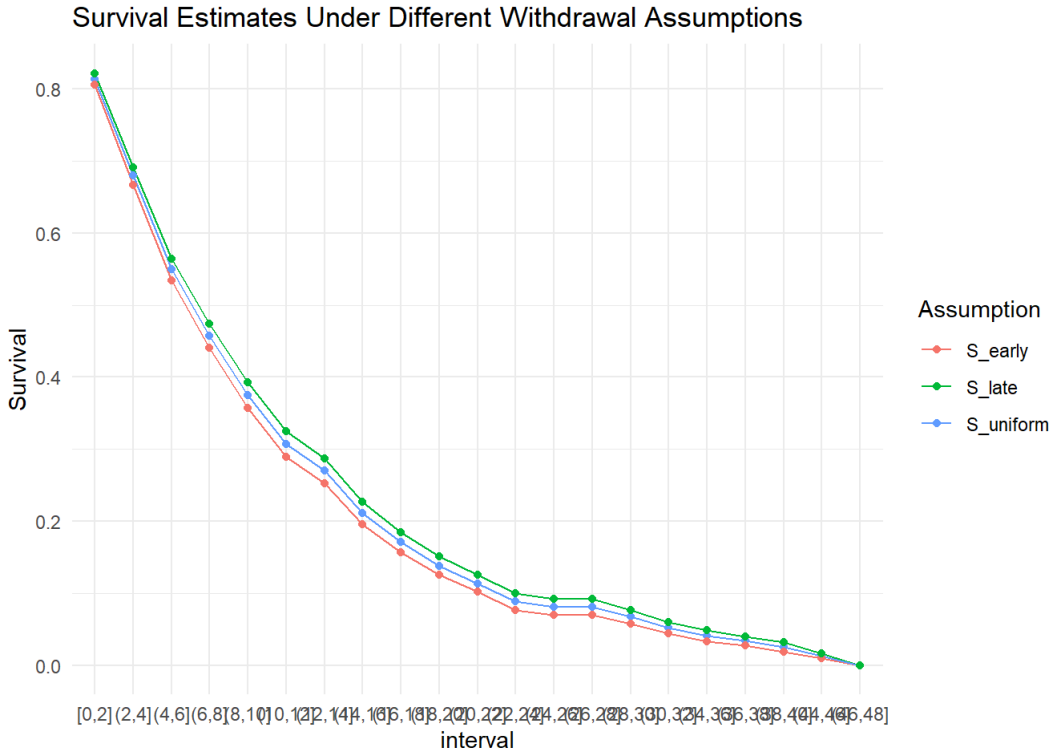


Findings: The plot compares True Survival with the Life Table Survival Estimate under the Uniform Withdrawal assumption. Both curves follow a similar decreasing pattern, indicating that the life table method provides a reasonably accurate approximation of the true survival function.Small deviations are visible, especially at later time points, where the life table estimate is slightly higher than the true survival.

Interpretation: The overall fit suggests that the uniform withdrawal assumption works well for this simulated dataset. Survival probability decreases sharply in the early time periods and gradually approaches zero as time increases.

1.5 Comparison of All Life-Table Assumptions

```
life_tab_long_sim <- lt_sim %>%
select(interval, S_uniform, S_early, S_late) %>%
pivot_longer(cols = c(S_uniform, S_early, S_late),
names_to = "Assumption",
values_to = "Survival")

ggplot(life_tab_long_sim, aes(x = interval, y = Survival, color = Assumption, group = Assumption)) +
geom_line() +
geom_point() +
theme_minimal() +
labs(title = "Survival Estimates Under Different Withdrawal Assumptions")
```



Findings: All three curves (early, late, uniform withdrawal) show a steady decline in survival probability as the interval increases, indicating increasing event occurrence over time.The late-withdrawal assumption gives the highest survival estimates, as it assumes individuals remain at risk for longer.The early-withdrawal assumption gives the lowest survival estimates, because individuals are assumed to leave earlier, reducing the number at risk.The uniform-withdrawal assumption lies between early and late, giving moderate survival probabilities.

Interpretation: All curves converge towards near-zero survival in the final intervals, suggesting most simulated subjects experience the event by the end of follow-up. Differences among the three assumptions are small, especially in late intervals, indicating that the choice of assumption has a limited impact on long-term survival estimates.

1.6 KM Estimator (Simulated Data)

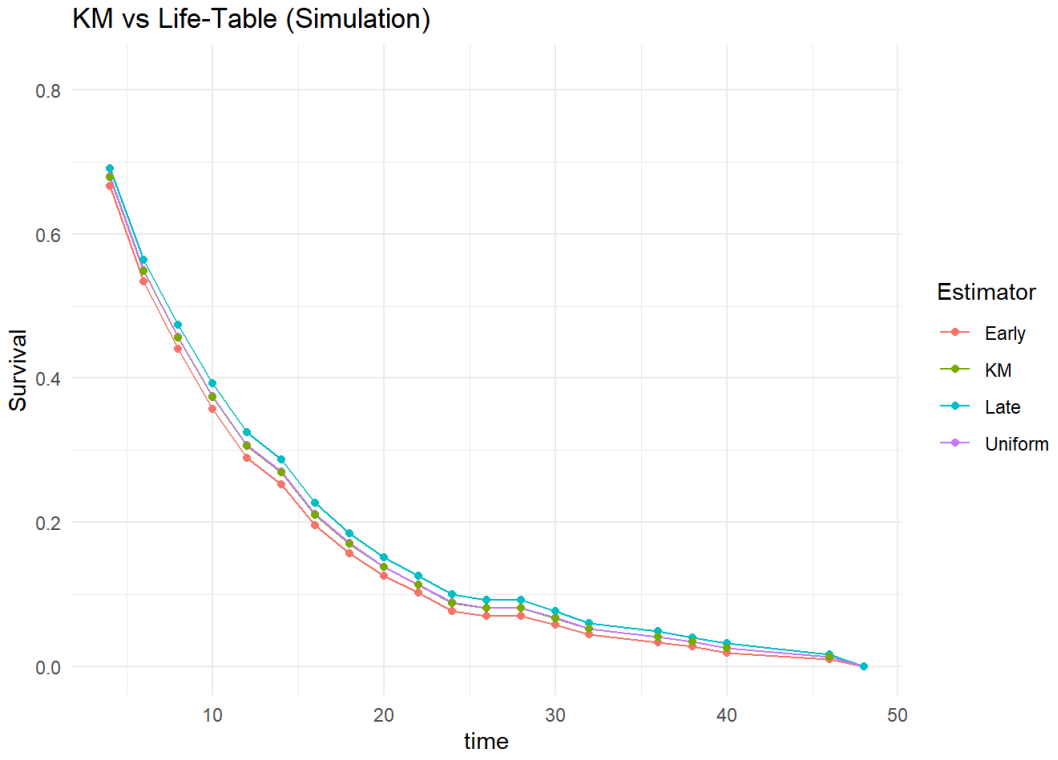
```
km_fit <- survfit(Surv(time, status) ~ 1, data = s_df)
km_summary <- summary(km_fit, times = breaks[-1])

km_df <- data.frame(
time = km_summary$time,
KM_survival = km_summary$surv
)
```

```
head(km_df)
```

```
##   time KM_survival
## 1    2   0.8124834
## 2    4   0.6784807
## 3    6   0.5489959
## 4    8   0.4567435
## 5   10   0.3741118
## 6   12   0.3061978
```

1.7 KM vs Life-Table Comparison

```
act_df <- data.frame(
time = interval_ends,
Uniform = lt_sim$S_uniform,
Early   = lt_sim$S_early,
Late    = lt_sim$S_late
)

compare_all <- left_join(act_df, km_df, by="time") %>%
rename(KM = KM_survival)

compare_all_long <- compare_all %>%
pivot_longer(cols = c(Uniform, Early, Late, KM),
names_to = "Estimator",
values_to = "Survival")

ggplot(compare_all_long, aes(x = time, y = Survival, color = Estimator)) +
geom_line() +
geom_point() +
theme_minimal() +
labs(title = "KM vs Life-Table (Simulation)")
```



Findings: At early times (0–10) all curves start at the same point and decline similarly. The KM estimate lies between the life-table estimates. Early-withdrawal assumption gives slightly lower survival, while late-withdrawal gives slightly higher survival. Around mid-times (10–25) small separation becomes visible. Uniform and KM lie in between and closely match each other. At later times (25–40) all four curves continue to decline smoothly. Differences between estimators narrow as the number of surviving individuals decreases. At the tail (40+) all methods converge to nearly the same survival probability, approaching zero due to very few subjects remaining at risk.

Interpretation: KM curve serves as a reference and stays close to the uniform-withdrawal life-table estimate. Early and late withdrawal assumptions respectively underestimate and overestimate survival but only slightly.

2. Real Dataset Study

The kidney dataset is a real-life clinical trial dataset that records catheter failure times for patients undergoing kidney treatment. It contains four key variables:

1. time – Number of days until catheter failure or censoring.
2. delta – Event indicator:

```
1 = catheter failure occurred

0 = observation is censored
```

3. type – Catheter insertion method:

```
1 = surgically inserted catheter

2 = percutaneous catheter
```

4. id – Patient identifier, used to track repeated catheter insertions for the same patient.

2.1 Kidney Dataset Life Table

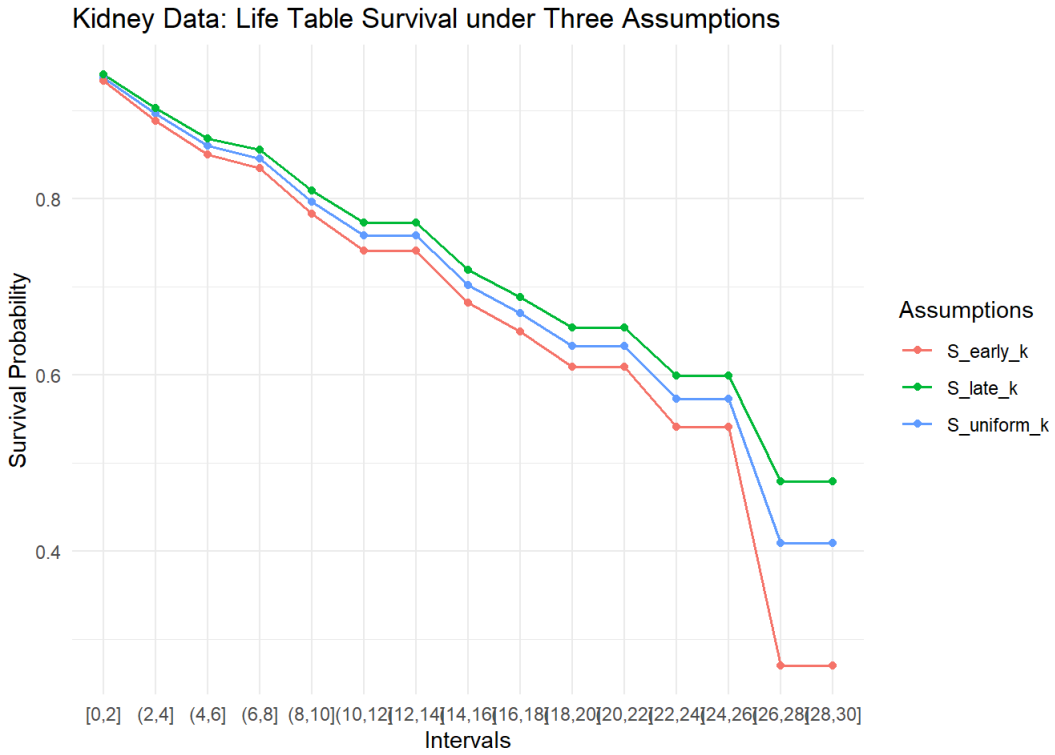
```
build_life_table_k <- function(df, breaks) {
df <- df %>% mutate(interval = cut(time, breaks = breaks, right = TRUE, include.lowest = TRUE))
lt <- df %>%
group_by(interval) %>%
summarise(
d_j_k = sum(delta == 1),
w_j_k = sum(delta == 0),
inside = n()
) %>% ungroup()

n_j_k <- numeric(nrow(lt))
n_j_k[1] <- nrow(df)
for(j in 2:nrow(lt)) {
n_j_k[j] <- n_j_k[j-1] - lt$d_j_k[j-1] - lt$w_j_k[j-1]
}

lt <- lt %>% mutate(
n_j_k = n_j_k,
q_uniform_k = d_j_k / (n_j_k - w_j_k/2),
q_early_k = d_j_k / (n_j_k - w_j_k),
q_late_k = d_j_k / n_j_k
)

lt <- lt %>% mutate(
S_uniform_k = cumprod(1 - replace_na(q_uniform_k, 0)),
S_early_k   = cumprod(1 - replace_na(q_early_k, 0)),
S_late_k    = cumprod(1 - replace_na(q_late_k, 0))
)

interval_ends <- as.numeric(gsub("\\((.+),(.+)\\]", "\\2", lt$interval))
lt$time <- interval_ends

return(lt)
}

data(kidney)
breaks_k <- seq(0, max(kidney$time) + 2, by = 2)
lt_k <- build_life_table_k(kidney, breaks_k)
```

```
head(lt_k)
```

```
## # A tibble: 6 × 12
##   interval d_j_k w_j_k inside n_j_k q_uniform_k q_early_k q_late_k S_uniform_k
##   <fct>    <int> <int>  <int> <dbl>       <dbl>     <dbl>    <dbl>       <dbl>
## 1 [0,2]        7    14     21   119      0.0625    0.0667   0.0588       0.938
## 2 (2,4]        4    15     19    98      0.0442    0.0482   0.0408       0.896
## 3 (4,6]        3    10     13    79      0.0405    0.0435   0.0380       0.860
## 4 (6,8]        1    10     11    66      0.0164    0.0179   0.0152       0.846
## 5 (8,10]       3     7     10    55      0.0583    0.0625   0.0545       0.796
## 6 (10,12]      2     7      9    45      0.0482    0.0526   0.0444       0.758
## # i 3 more variables: S_early_k <dbl>, S_late_k <dbl>, time <dbl>
```

2.2 Life Table Plots (Kidney)

```
life_tab_long_kidney <- lt_k %>%
select(interval, S_uniform_k, S_early_k, S_late_k) %>%
pivot_longer(cols = c(S_uniform_k, S_early_k, S_late_k),
names_to = "Assumptions",
values_to = "Survival")

ggplot(life_tab_long_kidney, aes(x = interval, y = Survival, color = Assumptions,group=Assumptions)) +
geom_line(size = 0.6) +
geom_point() +
theme_minimal() +
labs(
title = "Kidney Data: Life Table Survival under Three Assumptions",
x = "Intervals", y = "Survival Probability"
)
```



Findings: The survival probability shows a consistent decline across intervals under all three withdrawal assumptions. This pattern is broadly similar to the simulated dataset; however, in the kidney dataset, the differences between assumptions are minimal during the early intervals and become slightly more noticeable only in the later intervals. This widening reflects the reduced number at risk as time progresses.

Interpretation: In the kidney dataset, the life-table survival estimates behave as expected under different withdrawal assumptions.

• Early-withdrawal assumption leads to slightly lower survival estimates,

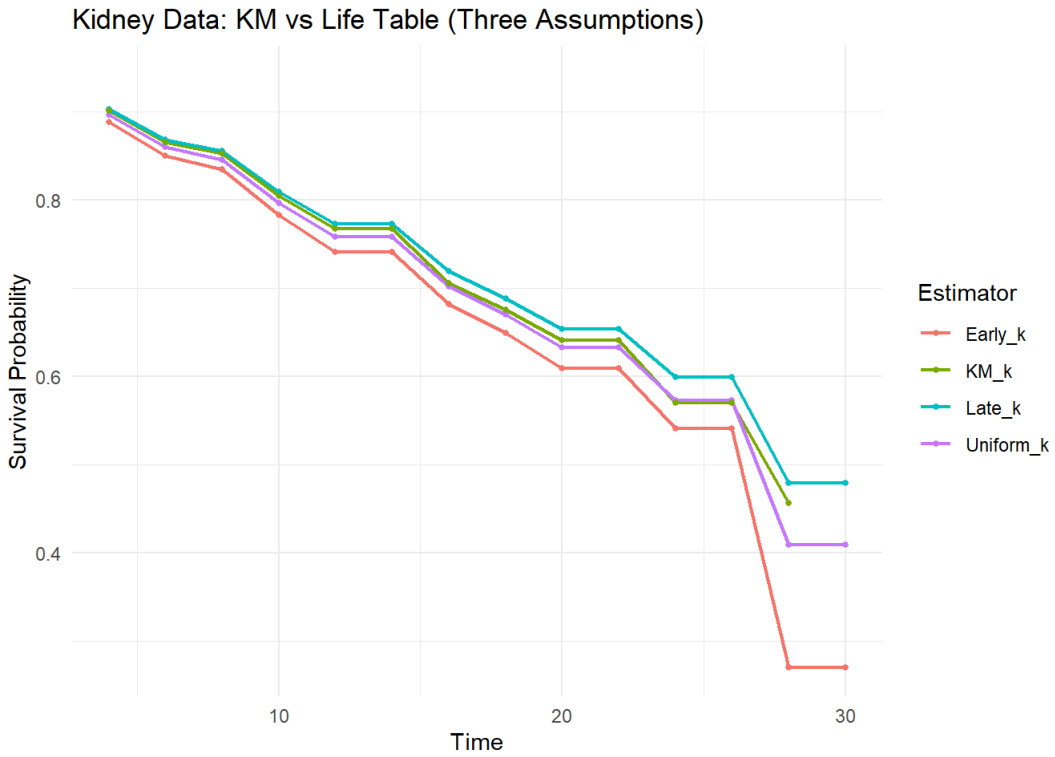
• Late-withdrawal results in marginally higher estimates, and

• Uniform-withdrawal lies between the two, closely following the overall Kaplan–Meier pattern.

The increasing separation between curves toward the end of follow-up is primarily due to the declining number of individuals at risk. When the sample size becomes small in later intervals, assumptions about censoring have a stronger influence on the estimated survival. This contrasts with the simulated dataset, where large sample size and ideal conditions produced much closer agreement among assumptions.

2.3 KM vs Life Table (Kidney)

```
km_fit_kidney <- survfit(Surv(time, delta) ~ 1, data = kidney)

km_summary_k <- summary(km_fit_kidney, times = breaks_k[-1])
km_df_k <- data.frame(time = km_summary_k$time, KM_survival_k = km_summary_k$surv)

act_df_k1 <- data.frame(
time = lt_k$time,
S_uniform = lt_k$S_uniform_k,
S_early   = lt_k$S_early_k,
S_late    = lt_k$S_late_k
)

compare_all_k <- left_join(act_df_k1, km_df_k, by = "time") %>%
rename(
Uniform_k = S_uniform,
Early_k  = S_early,
Late_k    = S_late,
KM_k      = KM_survival_k
)

compare_all_long_k <- compare_all_k %>%
pivot_longer(cols = c(Uniform_k, Early_k, Late_k, KM_k),
names_to = "Estimator",
values_to = "Survival")

ggplot(compare_all_long_k, aes(x = time, y = Survival, color = Estimator)) +
geom_line(size = 0.8) +
geom_point(size = 1) +
theme_minimal() +
labs(
title = "Kidney Data: KM vs Life Table (Three Assumptions)",
x = "Time",
y = "Survival Probability"
)
```



Findings: Overall pattern: All four estimators (KM and the three life-table versions) show a gradual decline in survival probability over time, as expected for the kidney dataset.

KM estimate (green): Serves as the benchmark. The life-table estimates stay close to this curve but differ depending on withdrawal assumptions.

Early-withdrawal assumption (red): Gives the lowest survival estimates throughout. Because censored individuals are assumed to withdraw early, more subjects are considered ‘at risk’, increasing the estimated event rate .

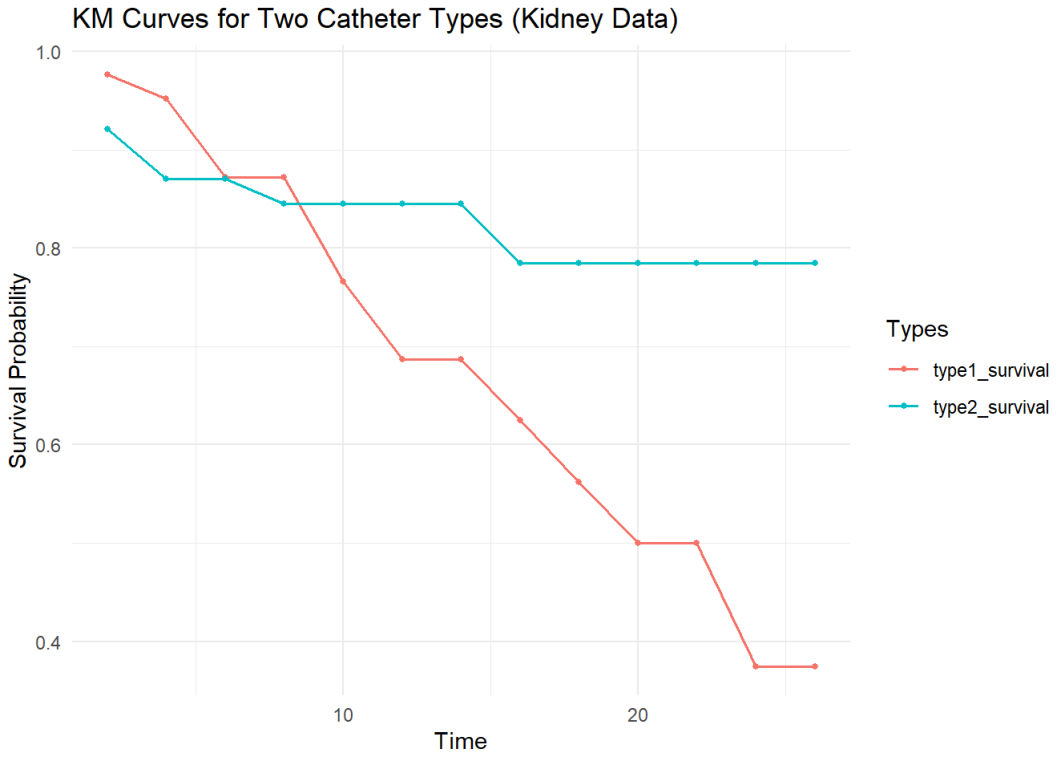
Late-withdrawal assumption (blue): Gives the highest survival estimates since censored individuals are assumed to withdraw late, so fewer are at risk indicating fewer effective failures resulting in higher survival.

Uniform-withdrawal assumption (purple): Lies between early and late, closely tracking the KM curve.It reflects the common assumption that censoring occurs uniformly within each interval. As the number at risk decreases, assumptions on withdrawals have a larger effect, causing wider separation between the curves.

Interpretation: The KM estimator remains stable, while life-table estimates vary slightly depending on censoring assumptions; early withdrawal underestimates survival most, late withdrawal overestimates, and uniform withdrawal approximates KM closely.

2.4 KM Curves by Catheter Type

```
fit1 <- survfit(Surv(time, delta) ~ 1, data = kidney[kidney$type==1,])
fit2 <- survfit(Surv(time, delta) ~ 1, data = kidney[kidney$type==2,])

fit1_summary <- summary(fit1, times = breaks_k[-1])
fit2_summary <- summary(fit2, times = breaks_k[-1])

fit1_df <- data.frame(time = fit1_summary$time, type1_survival = fit1_summary$surv)
fit2_df <- data.frame(time = fit2_summary$time, type2_survival = fit2_summary$surv)

com_df <- left_join(fit1_df, fit2_df, by = "time")

com_df_long <- com_df %>%
pivot_longer(cols = c(type1_survival, type2_survival),
names_to = "Types",
values_to = "Survival")

ggplot(com_df_long, aes(x = time, y = Survival, color = Types)) +
geom_line(size = 0.7) +
geom_point(size = 1) +
theme_minimal() +
labs(
title = "KM Curves for Two Catheter Types (Kidney Data)",
x = "Time",
y = "Survival Probability"
)
```



Findings: Type 2 catheter generally shows higher survival probabilities across the entire follow-up period compared to Type 1 whereas Type 1 catheter experiences more frequent and sharper drops, which indicates more failure events are occurring earlier and more often. During the early period (0–8 time units), survival of both types is similar, but Type 1 begins declining faster after this point.After around 10 time units, the curves separate clearly, with Type 2 maintaining stable survival while Type 1 continues to drop. And by the end of follow-up, Type 1 survival falls below 0.4, whereas Type 2 remains around 0.78, indicating a substantial difference.

Interpretation: Overall, the KM plot suggests that *Type 2 catheters have better survival performance than Type 1 catheters*.

Conclusion

In the simulated dataset, the uniform-withdrawal assumption closely approximates the true survival function because censoring was generated uniformly. In contrast, the kidney dataset shows larger deviations between life-table estimates and the KM curve. This is likely due to the smaller sample size and the unknown, non-uniform censoring pattern in the real data. Early-withdrawal underestimates survival, late-withdrawal overestimates it, and uniform-withdrawal lies in between, but does not match KM as closely as in the simulation.

Overall, the results suggest that in real datasets—especially those with moderate or smaller sample sizes like the kidney data—the choice of withdrawal assumption can introduce slight variations in survival estimates, particularly toward the tail of the distribution.

Additional Finding:
In the kidney dataset, Type 2 catheters showed better survival performance, indicating that they may be preferred for future use.

References:

• Bias in Classical Life Tables Under Censoring (https://www.researchgate.net/publication/396785857_Bias_in_Classical_Life_Tables_Under_Censoring_A_Comparative_Study_With_Kaplan-Meier_Estimation_and_Actuarial_Estimation_Using_Real_and_Simulated_Data)

• Book: Analysis of Failure and Survival Data : Book by P. Smith and Peter J Smith (https://api.pageplace.de/preview/DT0400.9781482295702_A31978986/preview-9781482295702_A31978986.pdf)