EAS509 Homework 6 (50 points).

Submit your answers as a single pdf attach all R code. Failure to do so will result in grade reduction.

Question 1 (25 points)

For each question, state whether or not the censoring mechanism is independent. Justify your answer with a short statement. (5 points for each)

a) In a study of disease relapse, due to a careless research scientist, all patients whose phone numbers begin with the number "2" are lost to follow up.

The censoring of patients in the study due to phone numbers starting with "2" appears random and not connected to the actual disease relapse. This suggests that such censoring is **INDEPENDENT** and does not impact the study's results.

b) In a study of longevity, a formatting error causes all patient ages that exceed 99 years to be lost (i.e. we know that those patients are more than 99 years old, but we do not know their exact ages).

The omission of precise age details for patients over 99 years old, caused by a formatting glitch, is a technicality and isn't related to the actual lifespan of the patients. This indicates that the censoring is **INDEPENDENT** of their longevity.

c) Hospital A conducts a study of longevity. However, very sick patients tend to be transferred to Hospital B, and are lost to follow up.

The transfer of patients to Hospital B because of critical illnesses suggests that the censoring is linked to the patients' health condition, a crucial factor in the longevity study. Therefore, the censoring is **NOT INDEPENDENT**.

d) In a study of unemployment duration, the people who find work earlier are less motivated to stay in touch with study investigators, and therefore are more likely to be lost to follow up.

Individuals finding work earlier are less motivated to stay in touch, suggesting that the duration of unemployment influences the likelihood of being lost to follow-up. This relationship between the study outcome (unemployment duration) and the likelihood of censoring indicates **NOT INDEPENDENT** censoring.

e) In a study of pregnancy duration, women who deliver their babies pre-term are more likely to do so away from their usual hospital, and thus are more likely to be censored, relative to women who deliver full-term babies.

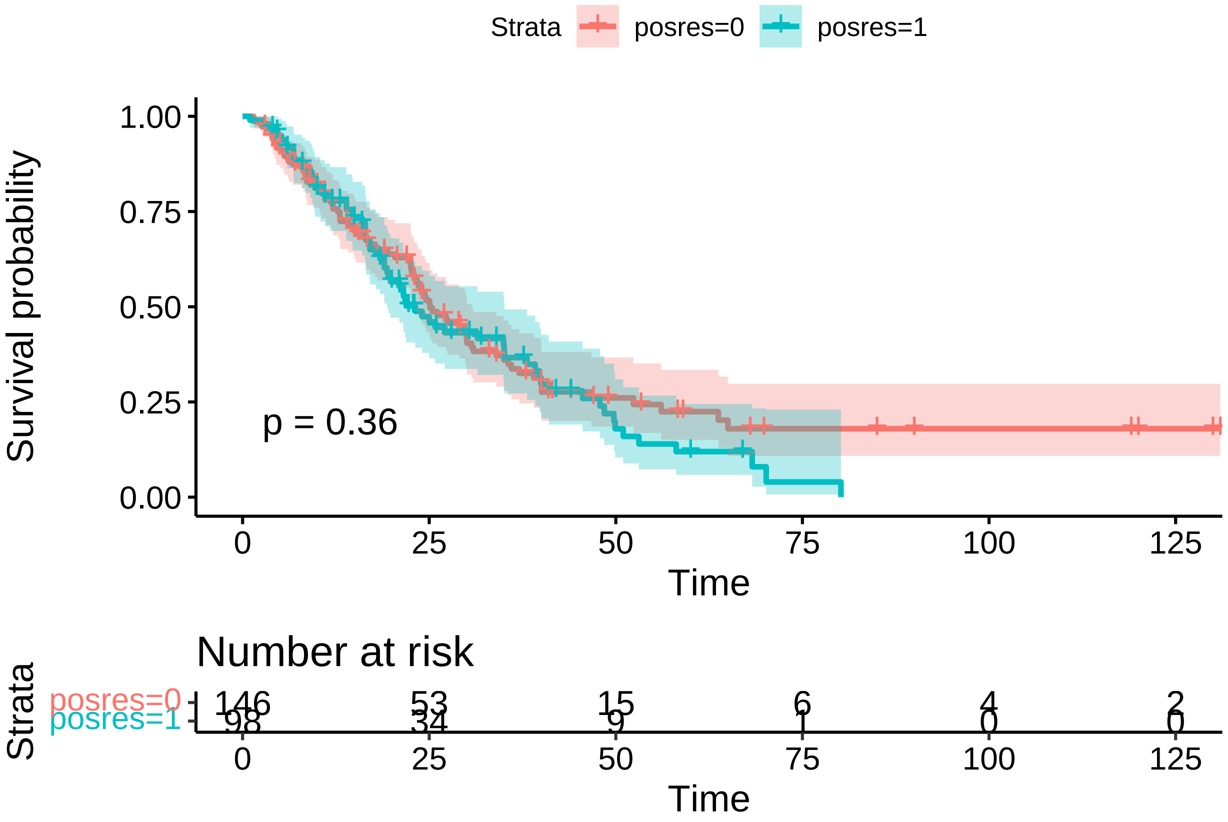
Women who give birth prematurely are more prone to being excluded from the study, which directly ties to the main focus of the research: the length of pregnancy. This indicates that the exclusion is affected by the key variable being examined, showing that the censoring is **NOT INDEPENDENT**.

Question 2 (25 points)

A data set from "DATA.csv" represents publication times for 244 clinical trials funded by the National Heart, Lung, and Blood Institute. Using Log-Rank Test in R, estimate if the Kaplan-Meier Survival Curves from two subpopulations stratified by "posres" variable are significantly different.

```
## Rows: 244 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): mech
## dbl (8): posres, multi, clinend, sampsize, budget, impact, time, status
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 244 x 9
##    posres multi clinend mech     sampsize budget impact  time status
##     <dbl> <dbl>   <dbl> <chr>       <dbl>  <dbl>  <dbl> <dbl>  <dbl>
##  1      0     0       1 R01         39876   8.02  44.0  11.2       1
##  2      0     0       1 R01         39876   8.02  23.5  15.2       1
##  3      0     0       1 R01          8171   7.61   8.39 24.4       1
##  4      0     0       1 Contract    24335  11.8   15.4   2.60      1
##  5      0     0       1 Contract    33357  76.5   16.8   8.61      1
##  6      0     0       1 Contract    10355   9.81  16.8   8.61      1
##  7      0     1       0 U01          1704  23.8    5.69 40.0       1
##  8      1     0       0 R01           150   2.70   3.50 27.1       1
##  9      0     0       0 R01           135   3.45   9.84 36.0       1
## 10      0     1       0 Contract      423  11.2   16.8   9.63      1
## # i 234 more rows
```



```
## Call:
## survdiff(formula = surv_obj ~ posres, data = data)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## posres=0 146       87     92.6     0.341     0.844
## posres=1  98       69     63.4     0.498     0.844
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

We can see a very high p-value which is much greater than alpha 0.05. So we fail to reject the Null-Hypothesis that there is no significant difference between the survival curves of the two groups.

That means there is no strong evidence to suggest that the "posres" variable significantly affects the survival distributions in your clinical trials dataset.