Nowacek HW 4-4

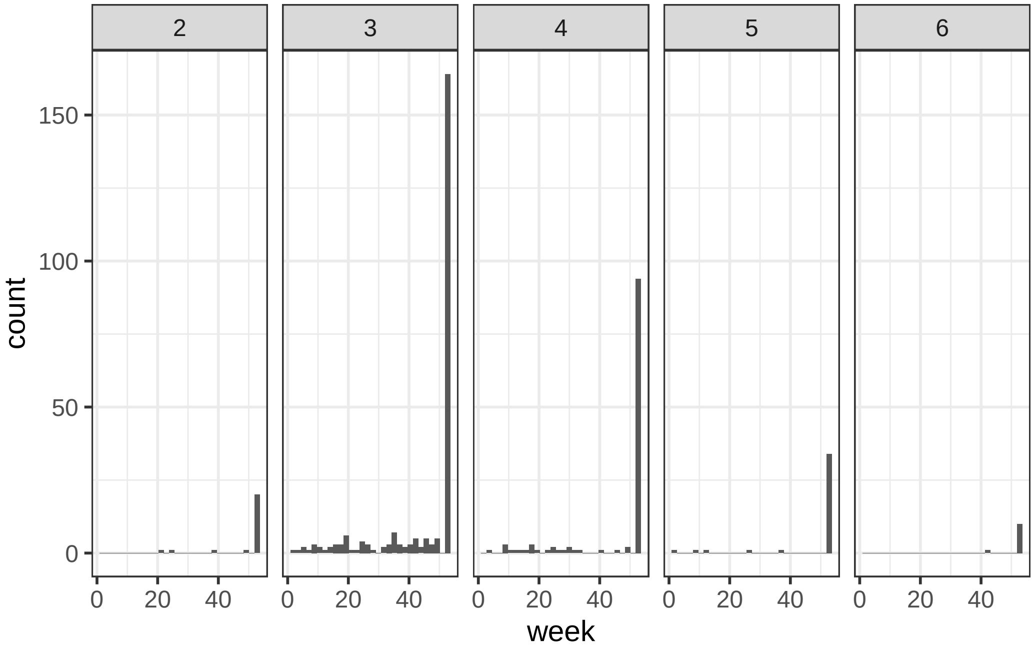
```r
d <- read_csv("https://www2.stat.duke.edu/courses/Fall24/sta490.01/data/hw4.csv")
```

```
Rows: 432 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (3): week, arrest, educ

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.
```

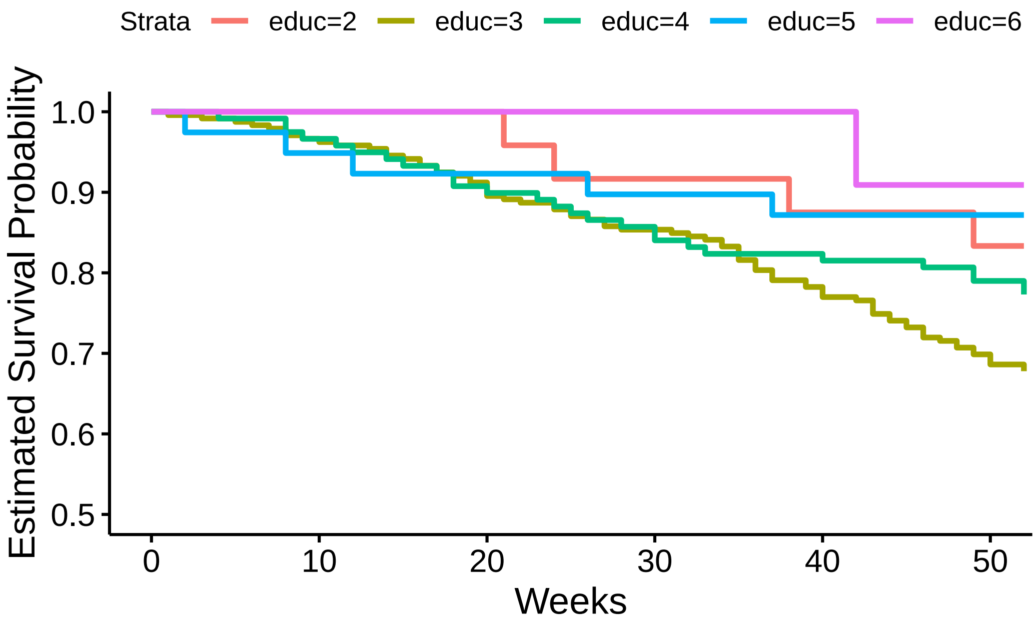
```r
ggplot(aes(x = week), data = d) +
  geom_histogram() +
  facet_grid(~ educ) +
  theme_bw()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
p <- survfit(Surv(week, arrest) ~ educ, data = d)

ggsurvplot(p,
           main = "Kaplan-Meier Estimates by Education Status",
           xlab = "Weeks",
           ylab = "Estimated Survival Probability",
           ylim = c(0.5, 1), conf.int = FALSE, censor = FALSE)
```



Here is the distribution of survival times, showing us the common censoring time, and the relative size of the groups. Followed by the survival curves for each group. Nothing in these graphs in incriminating for the log-rank test method, we now proceed with a log-rank test with the following hypotheses:

H_0: Under the null hypothesis, there is no significant association between education category and the time to recidivism in those groups. The survival functions across the education levels are assumed to be identical.

H_a: Under the alternate hypothesis, there is a significant association between education category and the time to recidivism in those groups. At least one of the survival functions across the education levels are assumed to be different than the rest.

```r
model <- survdiff(Surv(week, arrest) ~ educ, data = d)
model
```

```
Call:
survdiff(formula = Surv(week, arrest) ~ educ, data = d)

         N Observed Expected (O-E)^2/E (O-E)^2/V
educ=2  24        4     6.81     1.157      1.24
educ=3 239       77    61.54     3.884      8.49
educ=4 119       27    31.59     0.668      0.93
educ=5  39        5    10.78     3.102      3.45
educ=6  11        1     3.28     1.581      1.64
```

```
 Chisq= 10.5  on 4 degrees of freedom, p= 0.03
```

This model shows us that there are significant differences in the survival of groups based on education. The chi-square statistic in this case is 10.5, and on the chi-square distribution with four degrees of freedom (distribution under the null), the p-value is 0.03, indicating evidence for significant differences between the groups' survival at the $\alpha = 0.05$ level. This allows us to reject our null hypothesis that there are no effects of education on time to recidivism.