Salil-209-Project

Salil

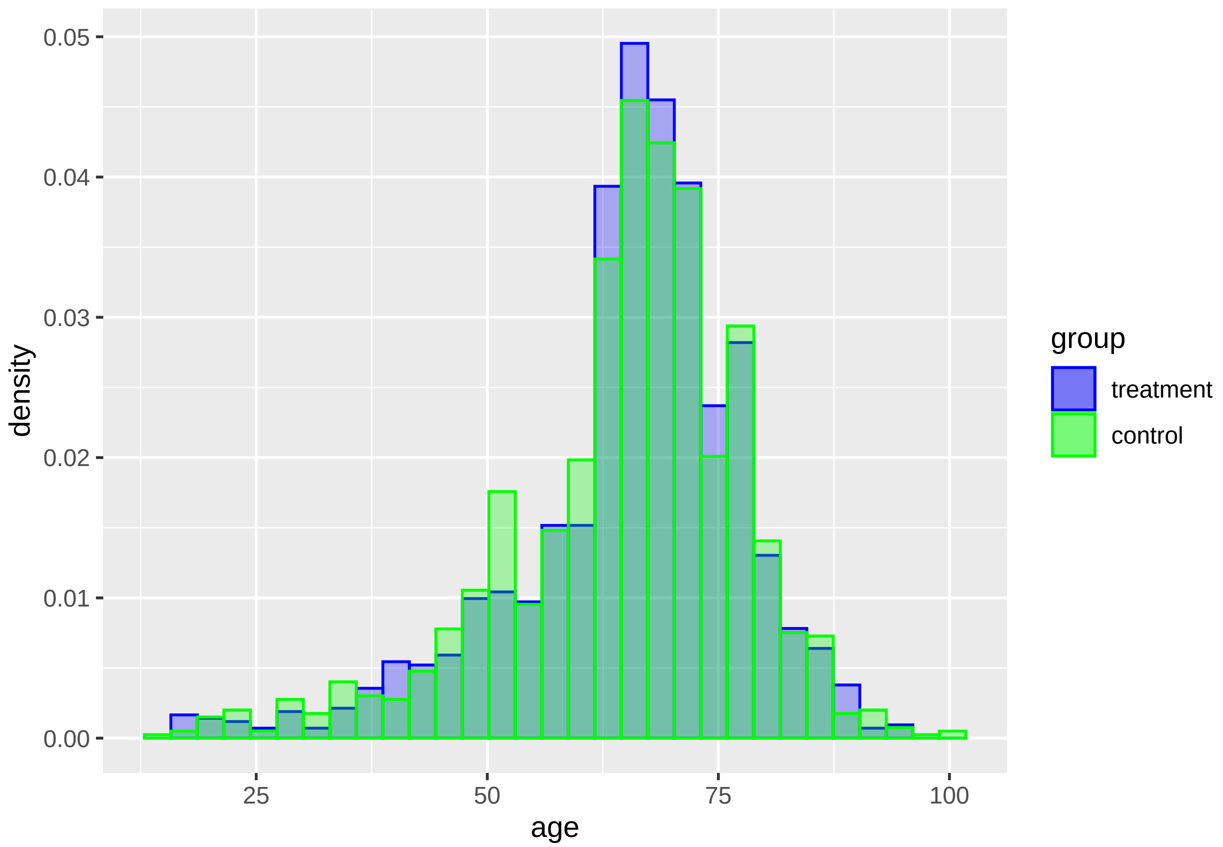
2023-12-02

```
library(ggplot2)
df <- read.table('fludata.txt')
```

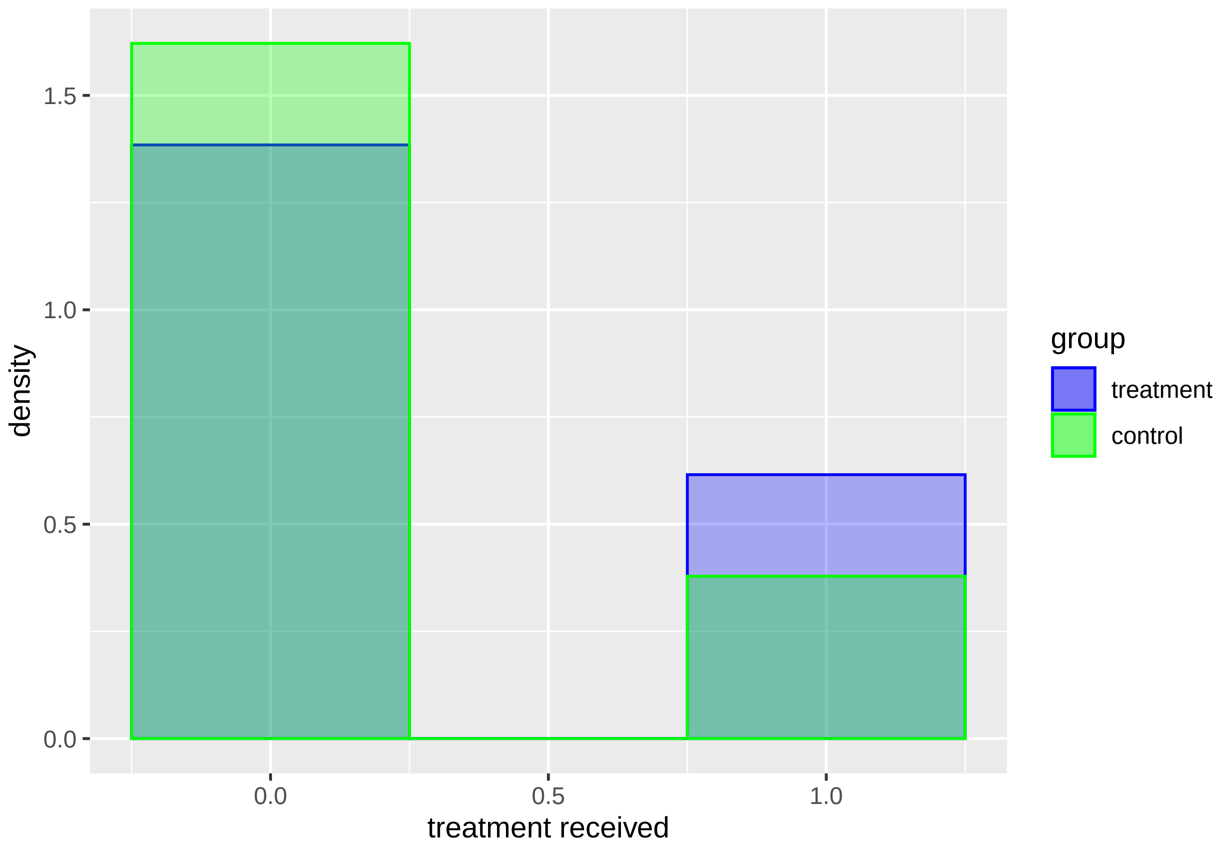
EDA

```
make_hists <- function(var, z, varname) {
  var_treat <- var[z==1]
  var_control <- var[z==0]

  # if the variable is an indicator, use binwidth 0.5
  if (length(unique(var))==2) {
    binwidth <- 0.5
  } else {
    binwidth <- (max(unique(var)) - min(unique(var)))/30
  }

  ggplot() +
    geom_histogram(aes(x=var_treat, y=..density.., col='b', fill='b'), alpha=0.3, binwidth=binwidth) +
    geom_histogram(aes(x=var_control, y=..density.., col='g', fill='g'), alpha=0.3, binwidth=binwidth) +
    scale_colour_manual(name="group", values=c("g" = "green", "b"="blue"), labels=c("g"="control", "b"=
    scale_fill_manual(name="group", values=c("g" = "green", "b"="blue"), labels=c("g"="control", "b"="t
    labs(x=varname)
}
```

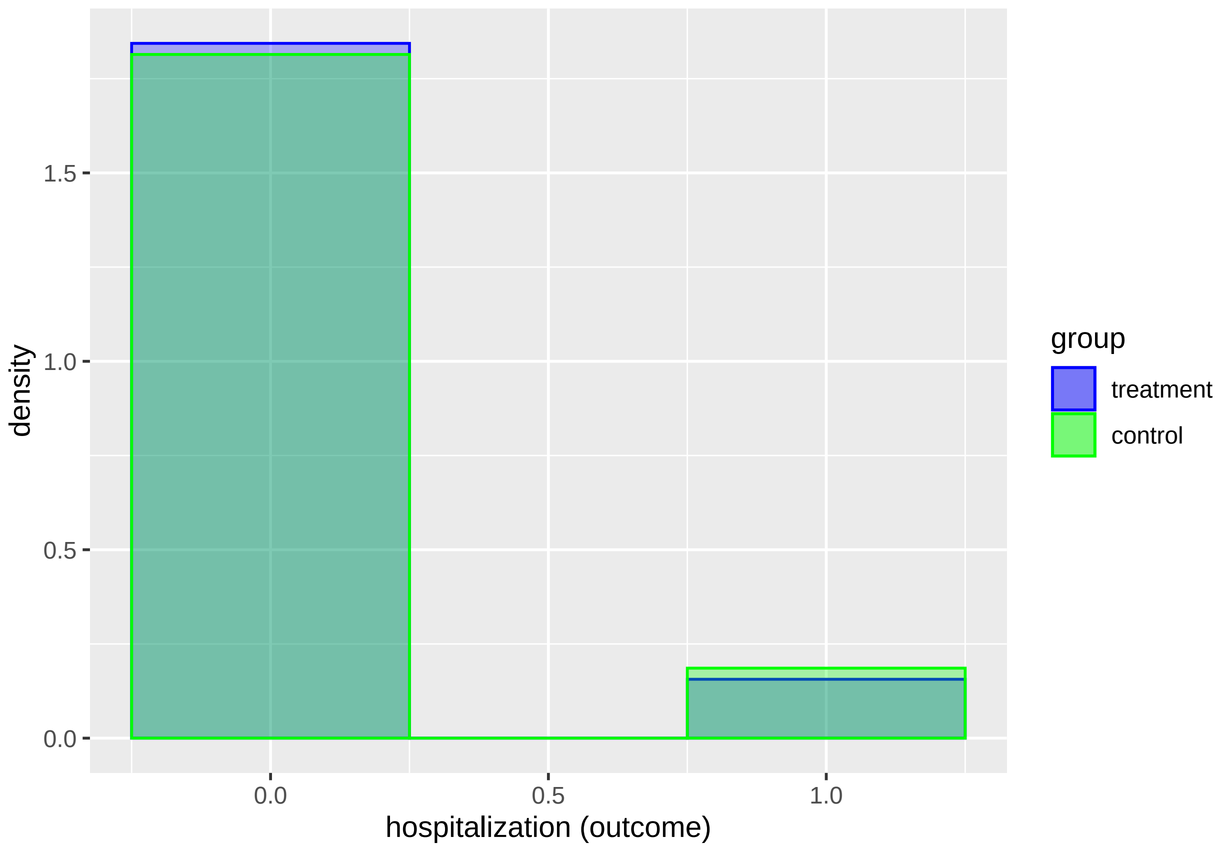
```
make_hists(df$age, df$assign, "age")
```



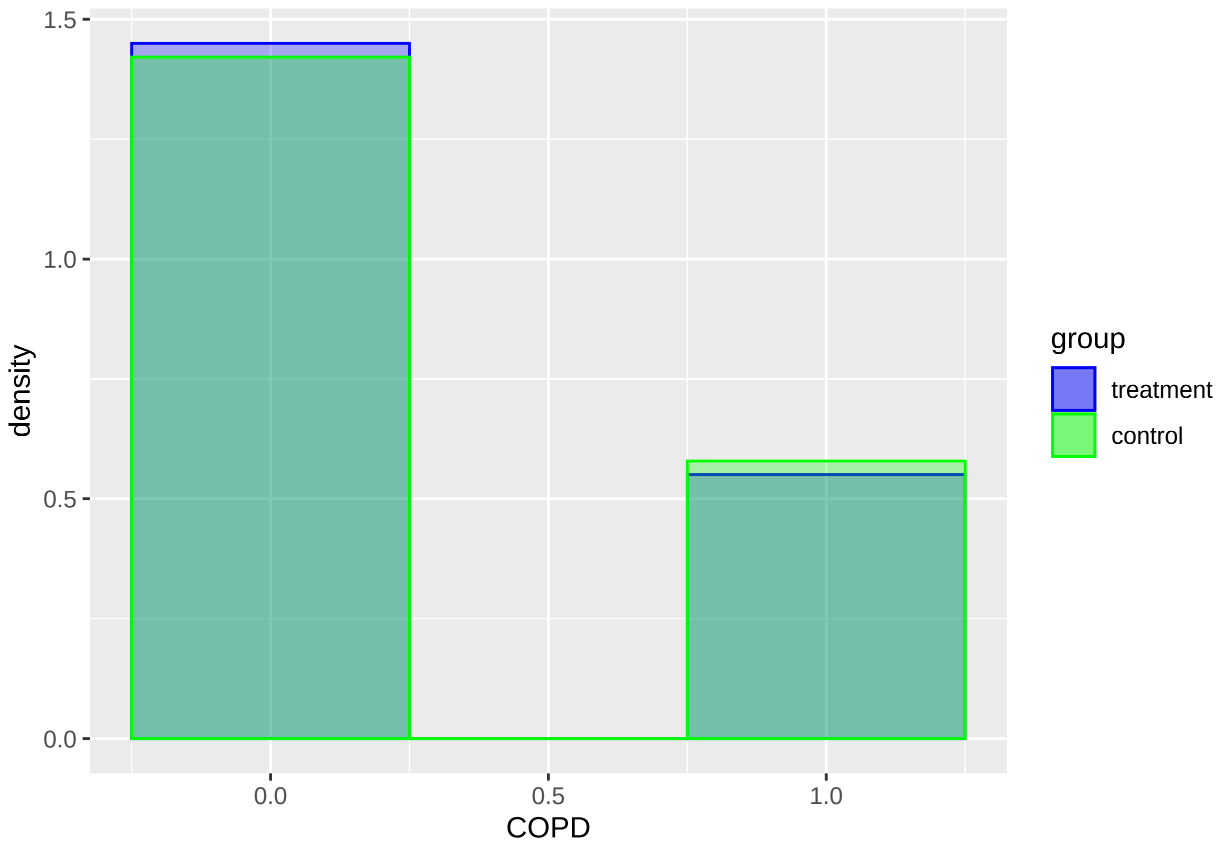
```
make_hists(df$receive, df$assign, "treatment received")
```



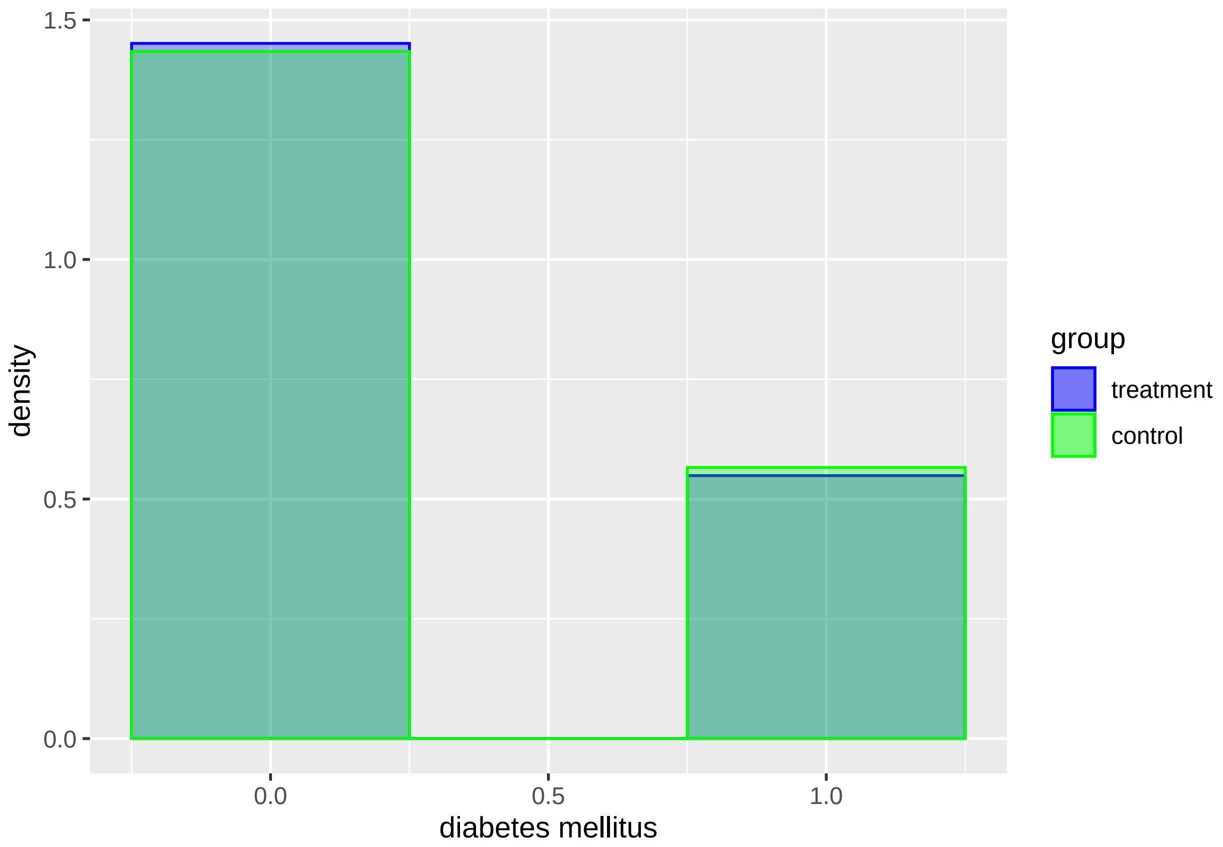
```r
make_hists(df$outcome, df$assign, "hospitalization (outcome)")
```



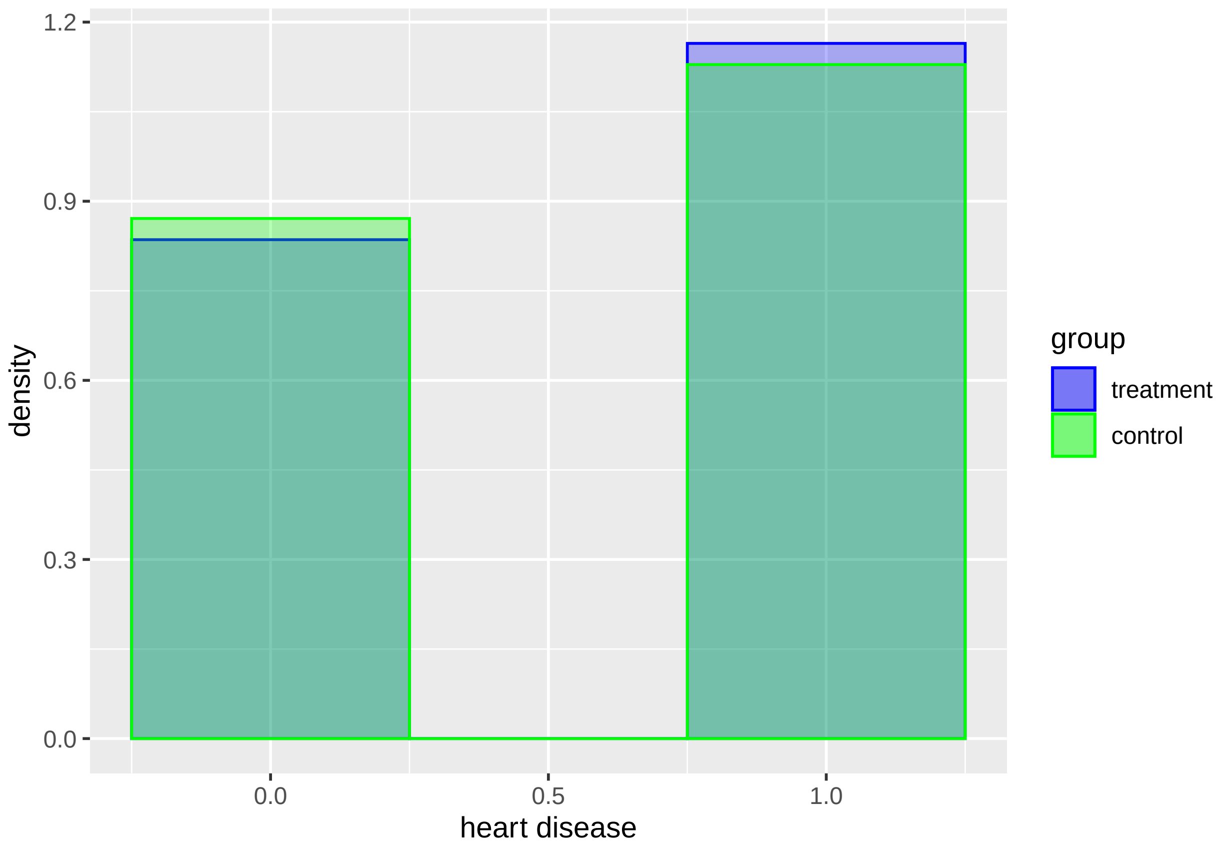
```
make_hists(df$copd, df$assign, "COPD")
```



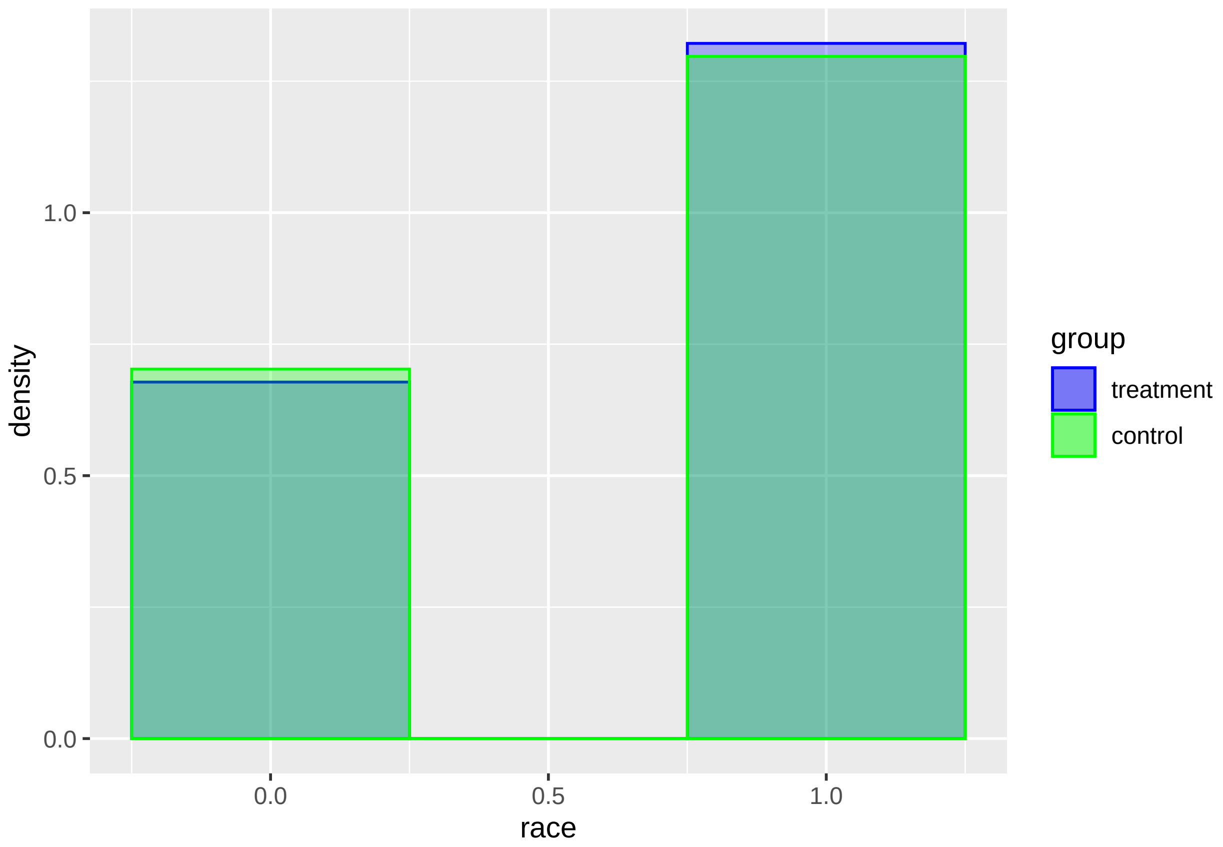
```
make_hists(df$dm, df$assign, "diabetes mellitus")
```



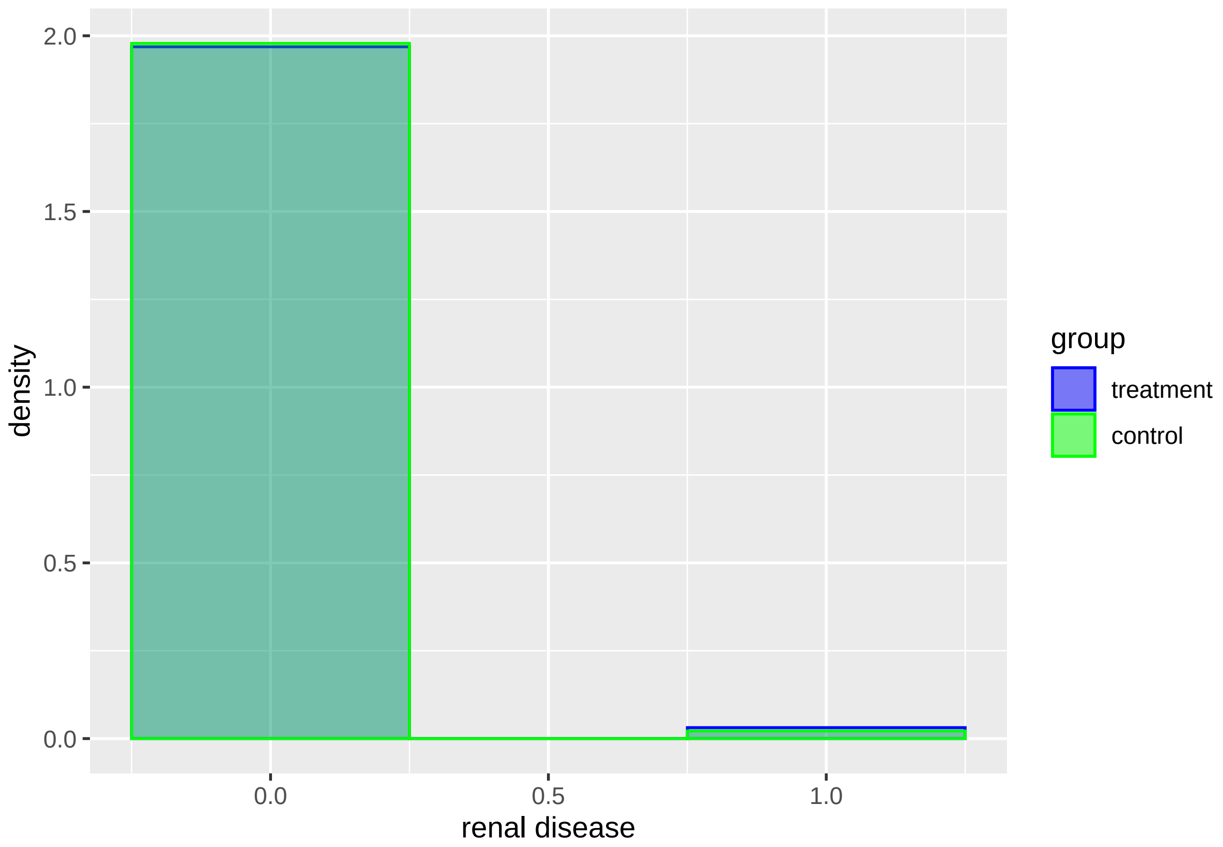
```r
make_hists(df$heartd, df$assign, "heart disease")
```



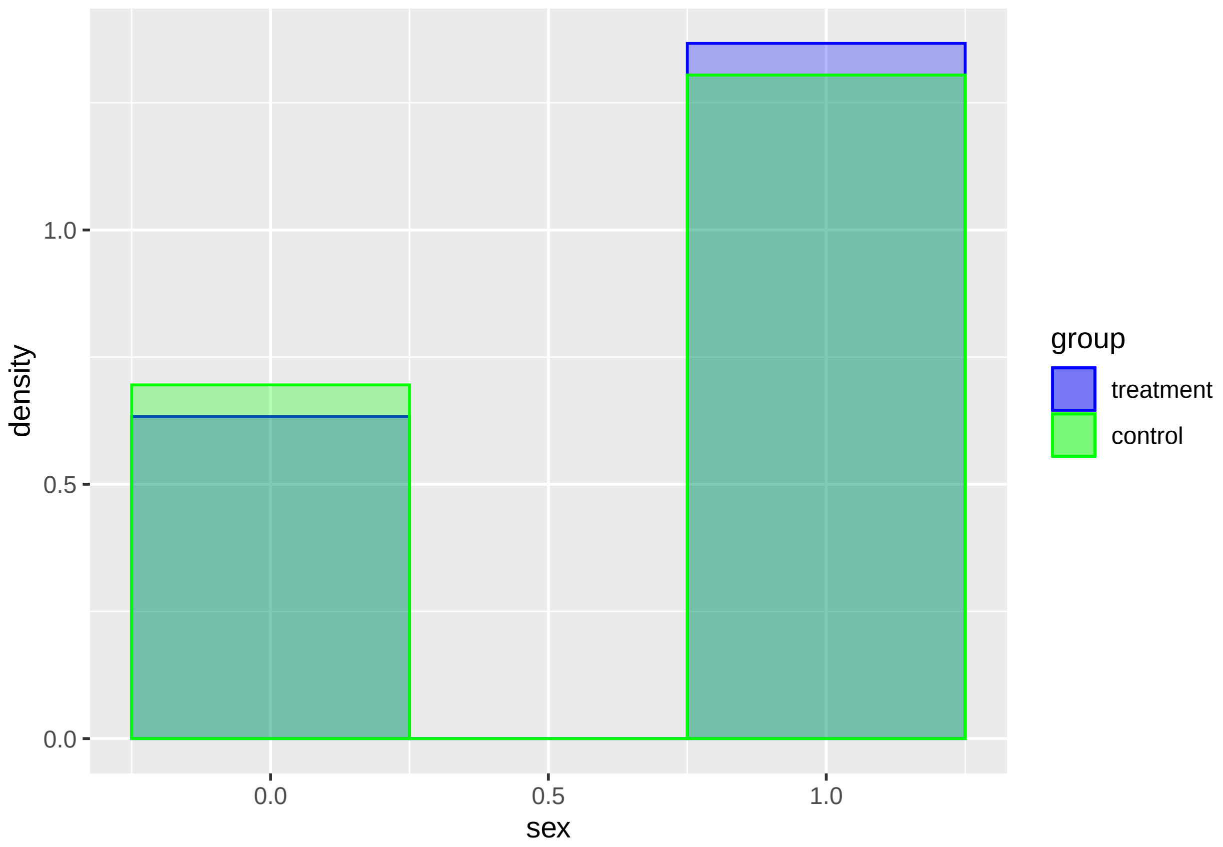
```
make_hists(df$race, df$assign, "race")
```



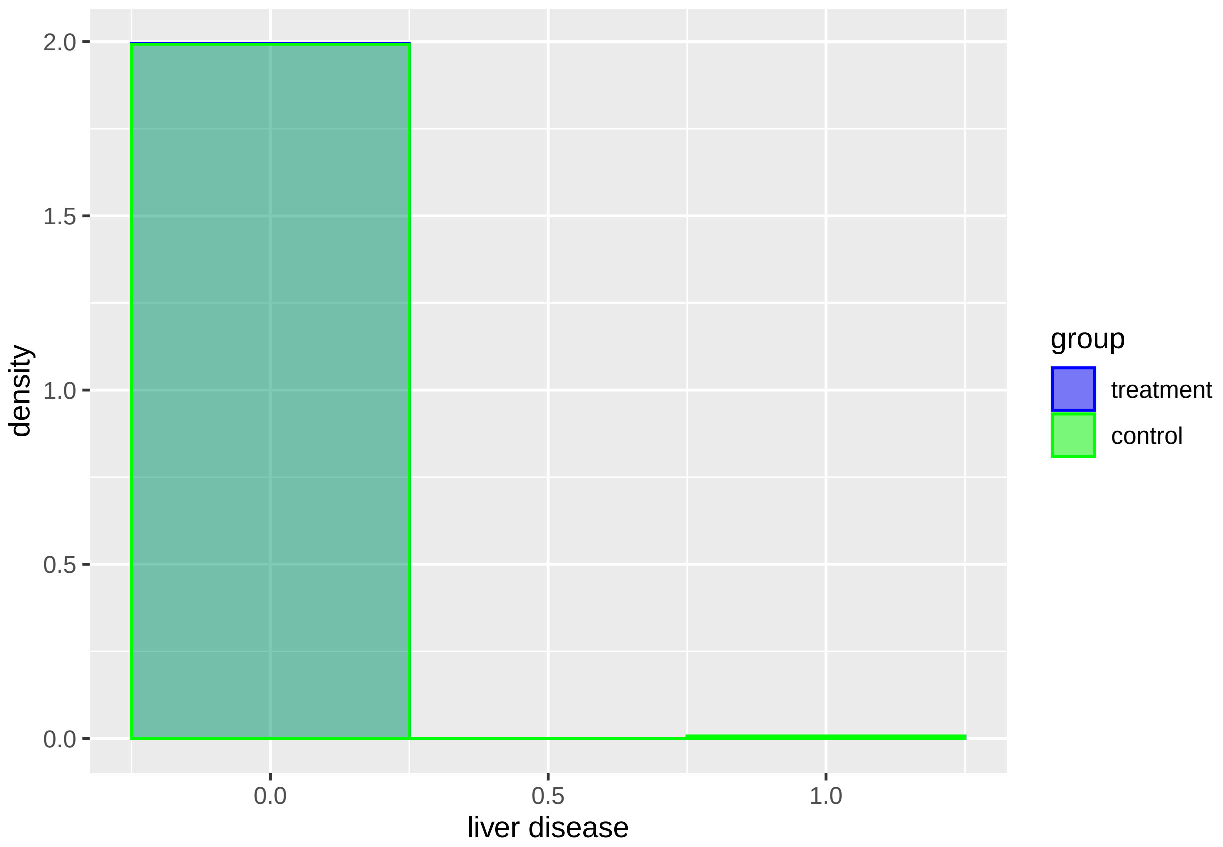
```
make_hists(df$renal, df$assign, "renal disease")
```



```
make_hists(df$sex, df$assign, "sex")
```



```r
make_hists(df$liverd, df$assign, "liver disease")
```



significance testing (ignore for now, just use t-test)

Most of the covariates are binary. So, to see whether their distributions are the same for treated and control units, we can do a test for difference in means of proportions.

Let's take `COPD` as the covariate. We want to test $H_0 : p_T = p_C$, where p_T is the proportion of treated units with COPD and p_C is the proportion of control units with COPD.

Since we have a large sample size, by the CLT, the difference in the proportion (i.e., sample average) of treated units and control that have COPD, $P_T - P_C$, is normally distributed with mean $p_T - p_C$ and variance $\frac{p_T q_T}{n_T} + \frac{p_C q_C}{n_C}$ (where $q_T = 1 - p_T$.)

Now, under H_0, . . . some stuff. Fill in later.

T-statistic is $T = \frac{\hat{p_T} - \hat{p_C} - 0}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}}$.

```
get_tstat <- function(var, z) {
  nT <- sum(z)
  nC <- length(z) - sum(z)

  var1z1 <- var[var==1 & z==1]
  var1z0 <- var[var==1 & z==0]
  var0z1 <- var[var==0 & z==1]
  var0z0 <- var[var==0 & z==0]

  phat <- (length(var1z1) + length(var1z0)) / length(var)
  pThat <- length(var1z1)/nT
  pChat <- length(var1z0)/nC

  return((pThat - pChat-0)/(phat * (1-phat) * ((1/nT) + (1/nC)) ))
}
```

```
get_tstat(df$copd, df$assign)
```

```
## [1] -50.39783
```

LOL something's wrong...