Exercise 2

Exercise introduction

This is Exercise 2 in Part 3 of the course.

The purpose of the exercise is to cover comparing groups. This will be done with t-tests and Wilcoxon Signed-rank test.

T-tests are valid when the assumptions of normal distribution and equal variance are met. If either assumption is violated, then a nonparametric alternative should be used (either Wilcoxon rank sum if unpaired or Wilcoxon signed-rank if paired data.)

Getting the data

Getting length & width data for female and male beetles.

```
(beetles <-
    data.frame(
        length = c(23, 24, 14, 15, 16, 12, 13, 9, 10, 14)
        , width = c(2, 3, 4, 3, 2, 4, 5, 5, 6, 6)
        , sex = c('Female', 'Female', 'Female', 'Female', 'Female', 'Male', 'Male', 'Male', 'Male', 'Ma
    ))
```

```
##    length width    sex
## 1      23     2 Female
## 2      24     3 Female
## 3      14     4 Female
## 4      15     3 Female
## 5      16     2 Female
## 6      12     4   Male
## 7      13     5   Male
## 8       9     5   Male
## 9      10     6   Male
## 10     14     6   Male
```

Calculating t-statistic

The t-statistic is just the mean difference divided by the standard error of difference.

```
# (aggBeetles <- beetles %>%
#     group_by(sex) %>%
#     summarize(
#         meanLength = mean(length)
#         , semLength = sd(length)/(sqrt(n()))
```

```
#         , meanWidth = mean(width)
#         , semWidth = sd(length)/(sqrt(n()))
#         )
# )
#
# (meanDiffLength <-
#     ((aggBeetles %>%
#         filter(sex == 'Female') %>%
#         select(meanLength)) -
#     (aggBeetles %>%
#         filter(sex == 'Male') %>%
#         select(meanLength))) %>%
#     rename(meanDiffLength = meanLength)
# )
#
# (seOfDiffLength <-
#     (sqrt(
#         (
#             aggBeetles %>%
#                 filter(sex == 'Female') %>%
#                 select(semLength)
#             ) ^ 2 +
#         (
#             aggBeetles %>%
#                 filter(sex == 'Male') %>%
#                 select(semLength)
#             ) ^ 2
#     )) %>% rename(seOfDiffLength = semLength)
#
#     )
#
# (tStatLength <- (meanDiffLength / seOfDiffLength) %>%
#         rename(tStatLength = meanDiffLength))
#
# (meanDiffWidth <-
#     ((aggBeetles %>%
#         filter(sex == 'Female') %>%
#         select(meanWidth)) -
#     (aggBeetles %>%
#         filter(sex == 'Male') %>%
#         select(meanWidth))) %>%
#     rename(meanDiffWidth = meanWidth)
# )
#
# (seOfDiffWidth <-
#     (sqrt(
#         (
#             aggBeetles %>%
#                 filter(sex == 'Female') %>%
#                 select(semWidth)
#             ) ^ 2 +
#         (
#             aggBeetles %>%
```

```
#                  filter(sex == 'Male') %>%
#                  select(semWidth)
#              ) ^ 2
#     )) %>% rename(seOfDiffWidth = semWidth)
#
#     )
#
# (tStatWidth <- (meanDiffWidth / seOfDiffWidth) %>%
#         rename(tStatWidth = meanDiffWidth))
```

Would then look up critical values for t-test based on the degrees of freedom. A two-sample t-test has degrees of freedom equal to (n_a - 1) + (n_b - 1). In this case, that is 8. A resource for t-critical values is here.

Considering that the second t-statistic I calculated was negative, I presume my calculations were wrong. If they are correct, though, and I presume I should just use the absolute value of the t-statistic, then there is a significant difference in length ~ sex at the 95% confidence level since the T_{CV} i 2.3060 and the t-statistic is 2.9482. However, there is no difference in width, as the t-statistic is 1.0405.

Also, when trying to knit this document without commenting out the code block above, an error is encountered where the variable `sex` is not found in the working scope. This is despite the code chunk working as desired when developing in RStudio. Thus, scoping is different while knitting. Furthermore, my use of the tidyverse syntax for doing the data manipulations in that code chunk were not very easy to follow, and there is probably a better way to do the intermediary calculations that would also be executable during the knitting process. For now, the code chunk is commented out since the insights from the code were not valuable and knitting the rest of the document is important.

Doing the t-tests with `Rcmdr`

```
library(Rcmdr)
```

```
## Warning: package 'Rcmdr' was built under R version 4.0.5

## Loading required package: splines

## Loading required package: RcmdrMisc

## Warning: package 'RcmdrMisc' was built under R version 4.0.5

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked _by_ '.GlobalEnv':
##
##     densityPlot
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: sandwich
```

```
## Loading required package: effects
```

```
## Warning: package 'effects' was built under R version 4.0.5
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
## The Commander GUI is launched only in interactive sessions
```

```
library(car)
library(RcmdrMisc)
library(rgl)
```

```
## Warning: package 'rgl' was built under R version 4.0.5
```

```
knitr::knit_hooks$set(webgl = hook_webgl)

t.test(length~sex, alternative='two.sided', conf.level=.95, var.equal=FALSE,
   data=beetles)
```

```
##
##  Welch Two Sample t-test
##
## data:  length by sex
## t = 2.9482, df = 5.4873, p-value = 0.02856
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.025844 12.574156
## sample estimates:
## mean in group Female   mean in group Male
##                 18.4                 11.6
```

```r
t.test(width~sex, alternative='two.sided', conf.level=.95, var.equal=FALSE,
   data=beetles)
```

```
##
##  Welch Two Sample t-test
##
## data:  width by sex
## t = -4.5356, df = 8, p-value = 0.00191
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.620223 -1.179777
## sample estimates:
## mean in group Female   mean in group Male
##                  2.8                  5.2
```

When using the `t.test` function, one can specify that one does not presume there to be equal variance. I presume this lowers the t-statistic by some empirically-derived method. The df for the width~sex model is *lower* than 8, however, which is surprising to me.
The conclusions of the analysis, however, are that female beetles are longer and narrower than male beetles.

Looking at differences between related groups

Will compare CO_2 emissions in 290 Swedish municipalities in 1990 and 2017.

```r
swedenCO2 <-
  read.table("../p02_inputs/CO2_municipalities.txt",
   header=TRUE, stringsAsFactors=TRUE, sep="\t", na.strings="NA", dec=",",
  strip.white=TRUE)
```

Do the paired t-test.

```r
with(swedenCO2, (t.test(X1990, X2017, alternative='two.sided',
  conf.level=.95, paired=TRUE)))
```

```
##
##  Paired t-test
##
## data:  X1990 and X2017
## t = 8.0182, df = 289, p-value = 0.00000000000002687
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  39253.42 64793.69
## sample estimates:
## mean of the differences
##                52023.55
```

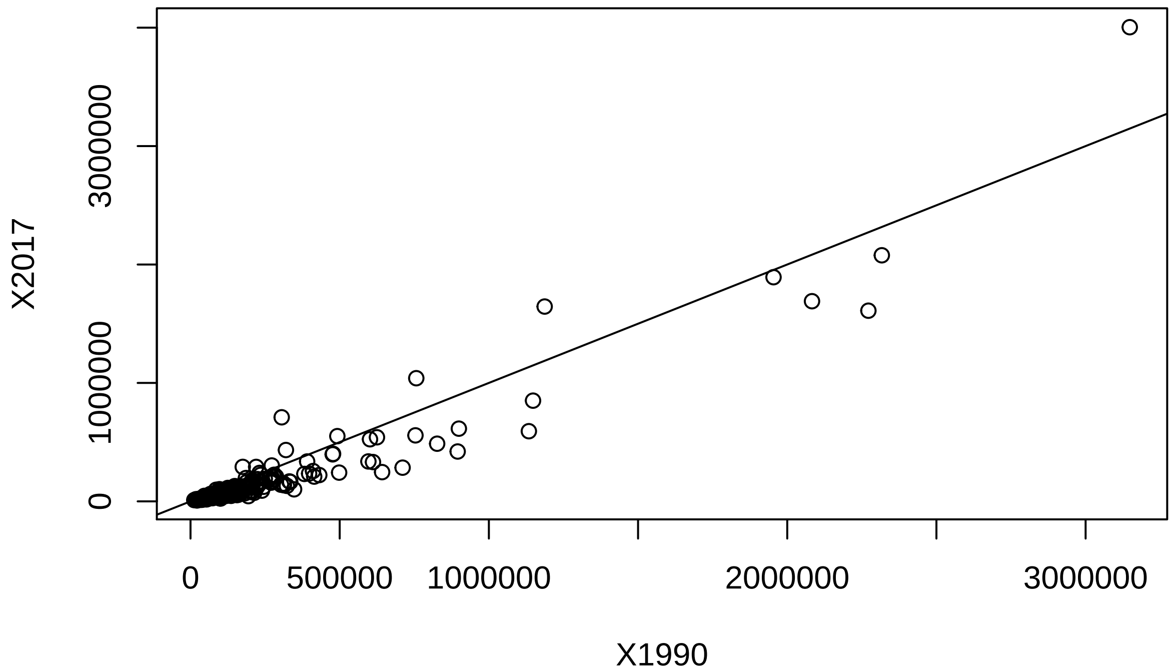
There is a significant difference, but the output of the `t.test` function doesn't tell which group is larger.

```
summary(swedenCO2)
```

```
##                    County         Municipality     X1990
##  VŠstra GŠtalands lŠn: 49   \200lmhult    :  1   Min.   :  12590
##  SkŒne lŠn           : 33   \200lvdalen   :  1   1st Qu.:  59540
##  Stockholms lŠn      : 26   \200lvkarleby:  1   Median : 100400
##  VŠrmlands lŠn       : 16   \200lvsbyn    :  1   Mean   : 197102
##  Dalarnas lŠn        : 15   \200ngelholm :  1   3rd Qu.: 187125
##  VŠsterbottens lŠn   : 15   ...ckerš     :  1   Max.   :3148000
##  (Other)             :136   (Other)   :284
##      X2000             X2005             X2010             X2011
##  Min.   :  14480   Min.   :  12280   Min.   :  12180   Min.   :  10510
##  1st Qu.:  56028   1st Qu.:  49368   1st Qu.:  43760   1st Qu.:  41905
##  Median :  93640   Median :  84755   Median :  80000   Median :  75380
##  Mean   : 187884   Mean   : 184980   Mean   : 181938   Mean   : 168362
##  3rd Qu.: 171925   3rd Qu.: 155450   3rd Qu.: 143225   3rd Qu.: 132200
##  Max.   :3312000   Max.   :3928000   Max.   :3863000   Max.   :3670000
##
##      X2012             X2013             X2014             X2015
##  Min.   :   9018   Min.   :   9038   Min.   :   8945   Min.   :   8691
##  1st Qu.:  40050   1st Qu.:  38670   1st Qu.:  37638   1st Qu.:  37675
##  Median :  70490   Median :  69255   Median :  64975   Median :  65305
##  Mean   : 159670   Mean   : 154195   Mean   : 148814   Mean   : 148511
##  3rd Qu.: 128550   3rd Qu.: 121700   3rd Qu.: 117275   3rd Qu.: 113850
##  Max.   :3383000   Max.   :3543000   Max.   :3603000   Max.   :2913000
##
##      X2016             X2017
##  Min.   :   8366   Min.   :   7650
##  1st Qu.:  36012   1st Qu.:  35712
##  Median :  63455   Median :  63140
##  Mean   : 146886   Mean   : 145079
##  3rd Qu.: 108250   3rd Qu.: 111175
##  Max.   :3878000   Max.   :4004000
##
```

The CO_2 emissions in Sweden has actually *decreased* from 1990 to 2017.

```
swedenCO2 %>%
    select(X1990, X2017) %>%
    plot()

abline(a = 0, b = 1)
```



This plot shows the data in the counties in 1990 and 2017 with the ‘line-of-identity’. All dots that lay below the line of identity had lower emissions in 2017 than they did in 1990.

Redoing the test with non-parametric Wilcoxon Signed-Rank

```
with(swedenCO2, median(X1990 - X2017, na.rm=TRUE)) # median difference
```

```
## [1] 33520
```

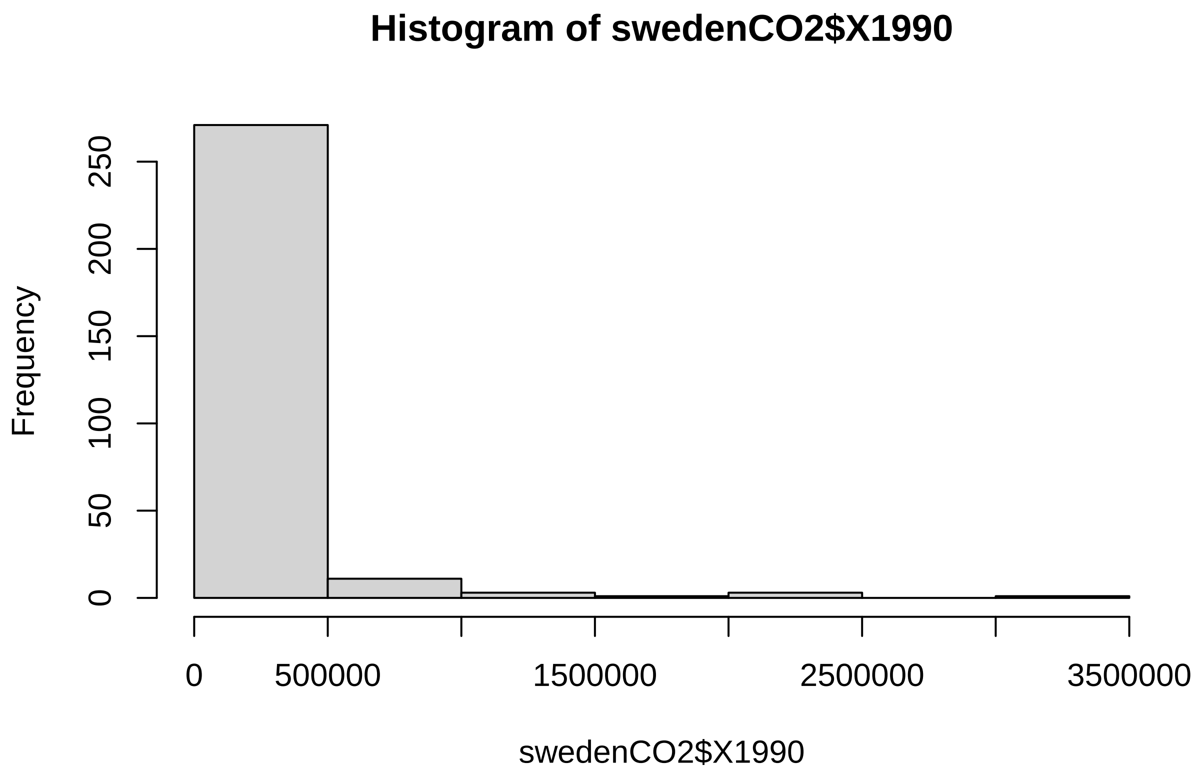
```
with(swedenCO2, wilcox.test(X1990, X2017, alternative='two.sided',
  paired=TRUE))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  X1990 and X2017
## V = 39940, p-value < 0.00000000000000022
## alternative hypothesis: true location shift is not equal to 0
```

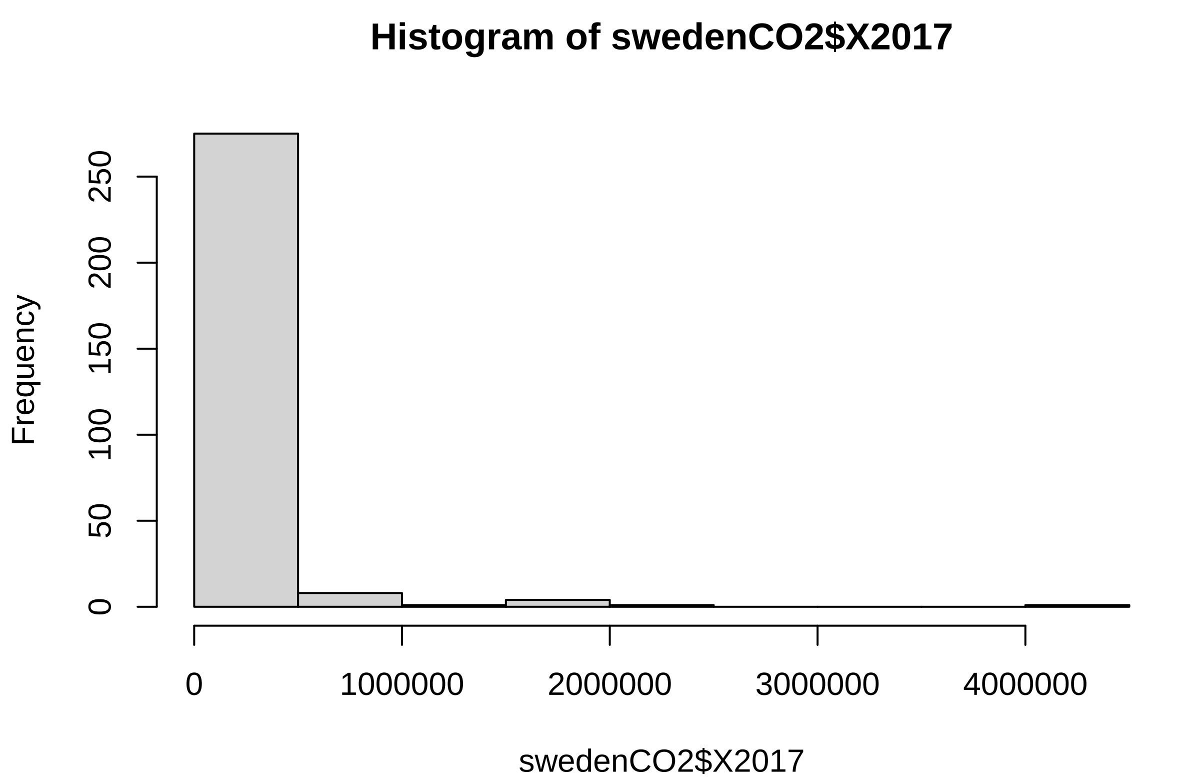
The nonparametric test also detected the significant difference, and the p-value was 1-2 orders of magnitude smaller. Which test should have been chosen based on ‘eyeballing’ the histograms of the years?

Histograms of years

```
hist(swedenCO2$X1990)
```

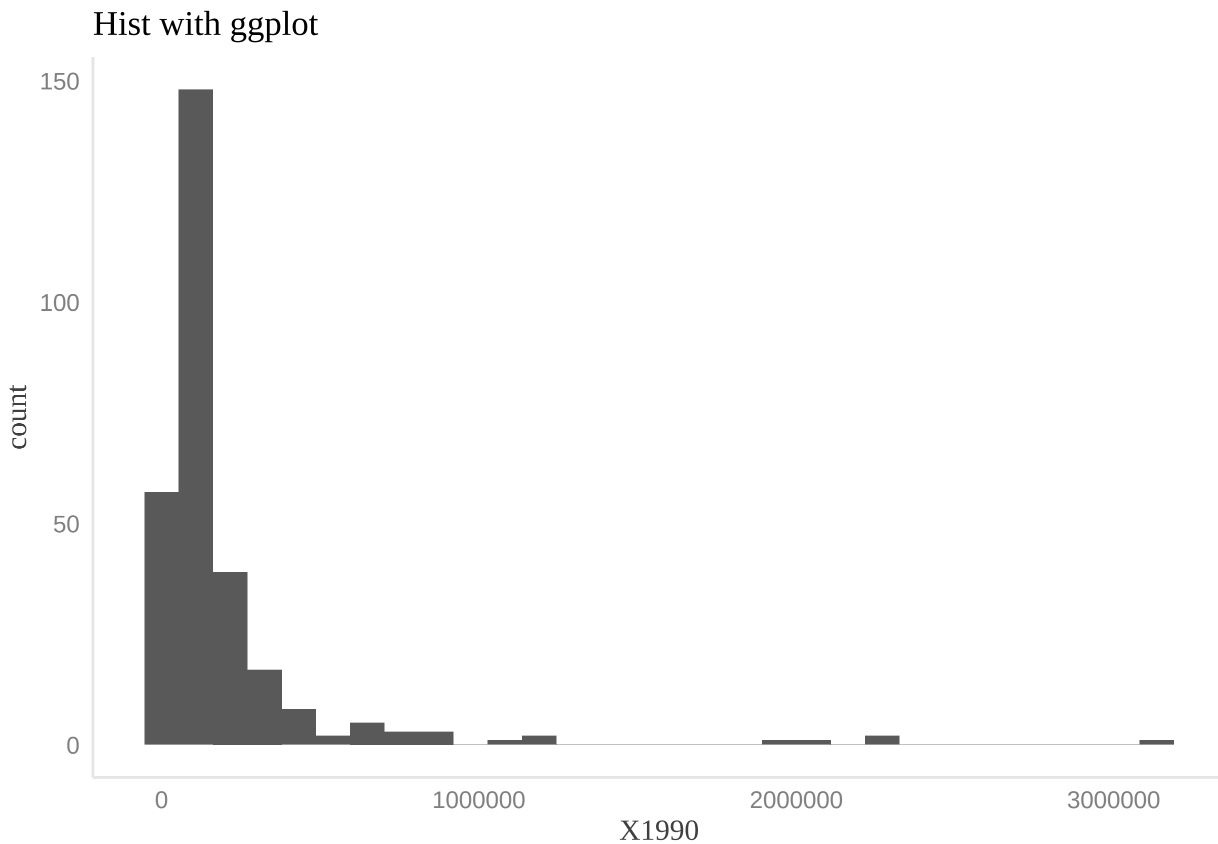



```
hist(swedenCO2$X2017)
```



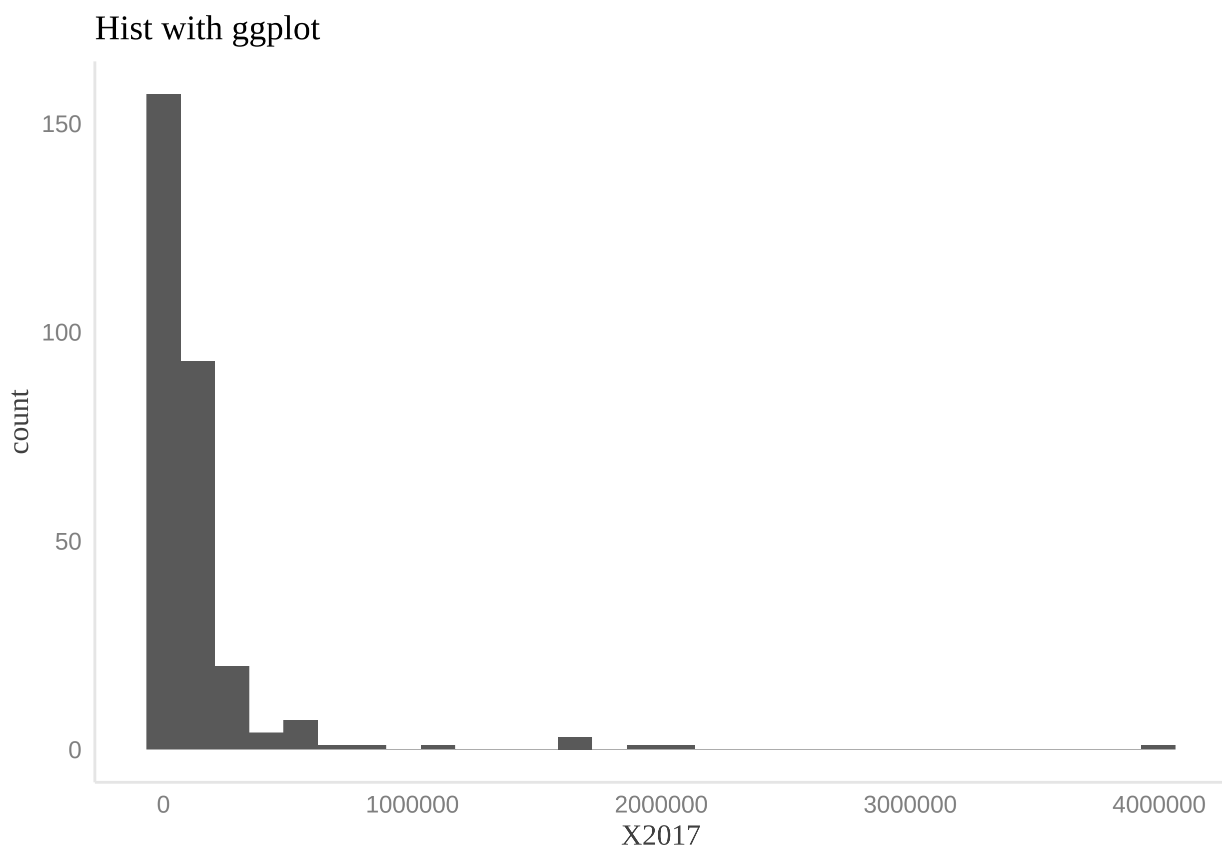
```r
ggplot(swedenCO2, aes(x = X1990)) +
    geom_histogram() +
    ggtitle('Hist with ggplot')
```

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




```r
ggplot(swedenCO2, aes(x = X2017)) +
    geom_histogram() +
    ggtitle('Hist with ggplot')
```

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

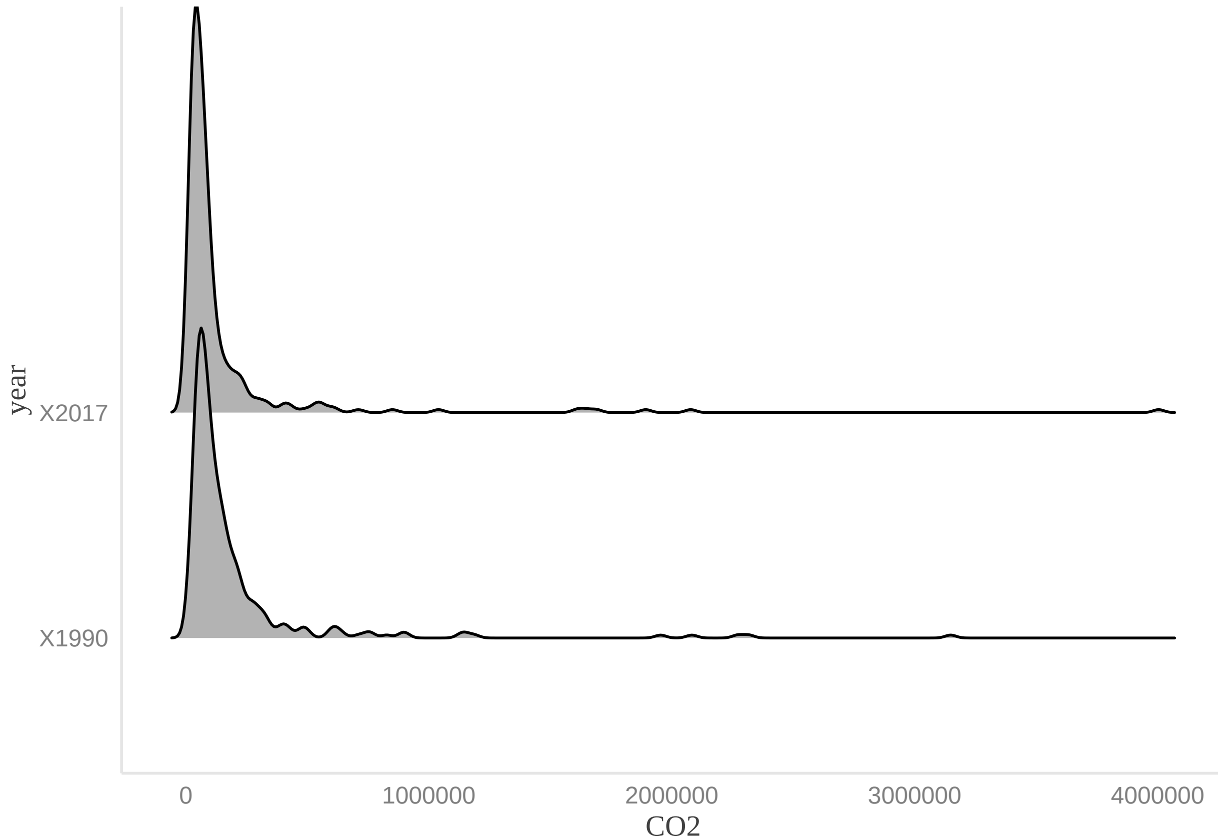



```
library(ggridges)
```

```
## Warning: package 'ggridges' was built under R version 4.0.5
```

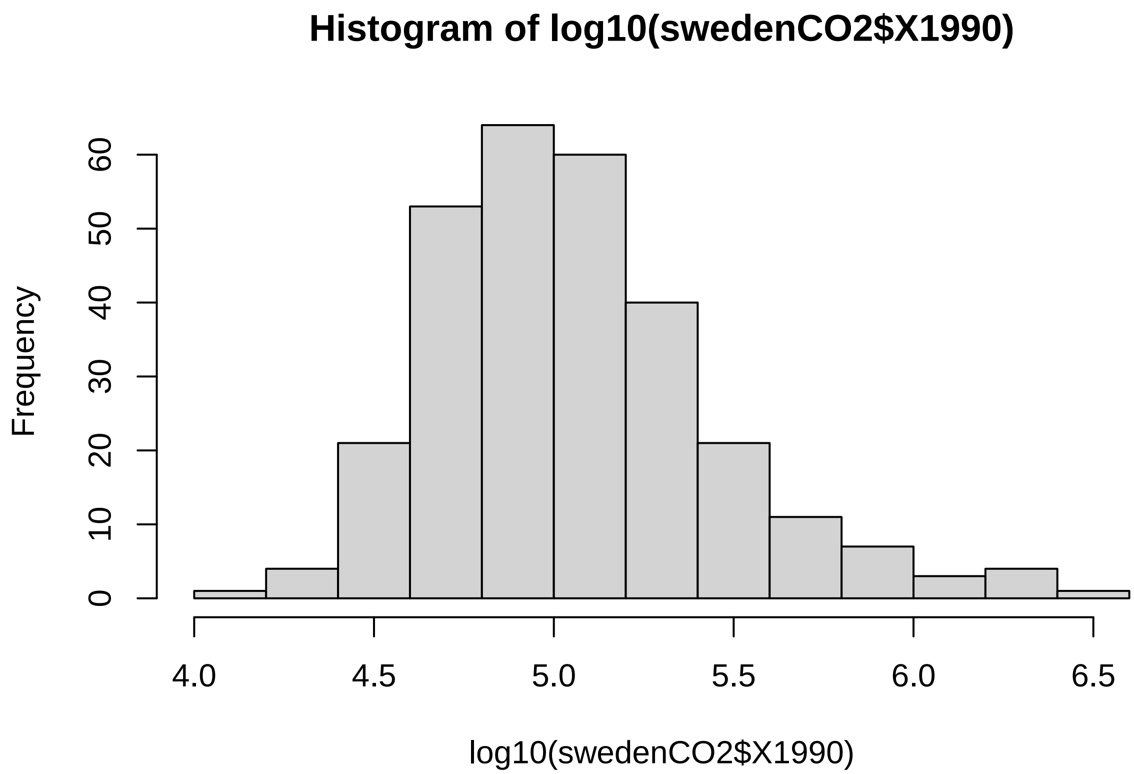
```
swedenCO2 %>%
    select(X1990,X2017) %>%
    pivot_longer(cols = c(X1990, X2017), names_to = 'year', values_to = 'CO2') %>%
    ggplot(aes(x = CO2, y = year)) +
    geom_density_ridges()
```

```
## Picking joint bandwidth of 21900
```

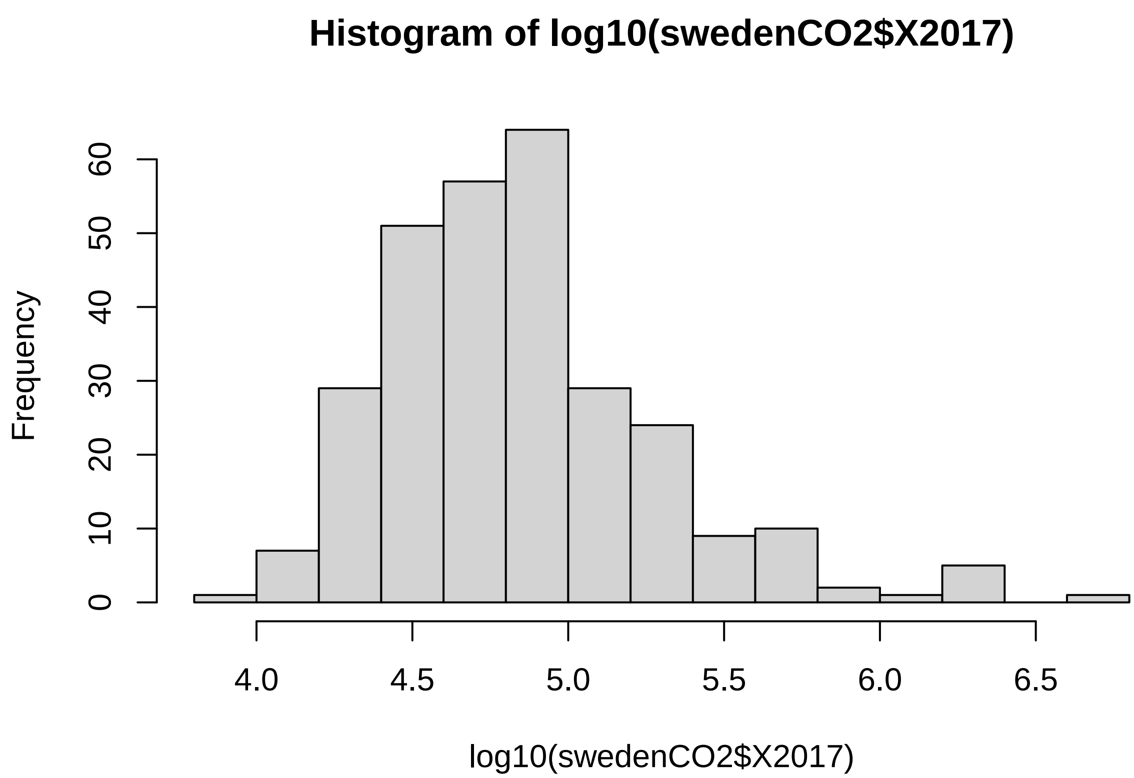


Neither dataset is anything close to being normally distributed. Either the data need to be transformed or non-parametric alternatives should be used.

```
hist(log10(swedenCO2$X1990))
```

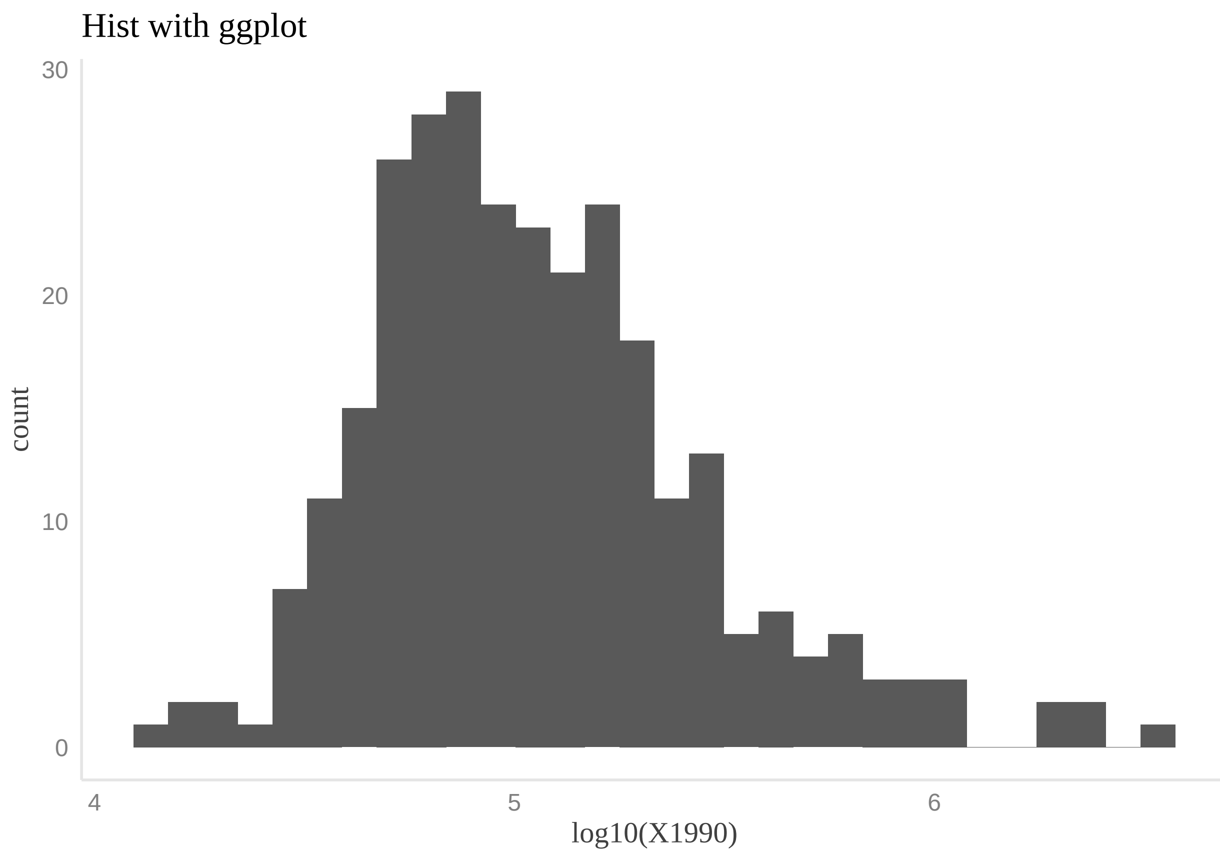



```
hist(log10(swedenCO2$X2017))
```



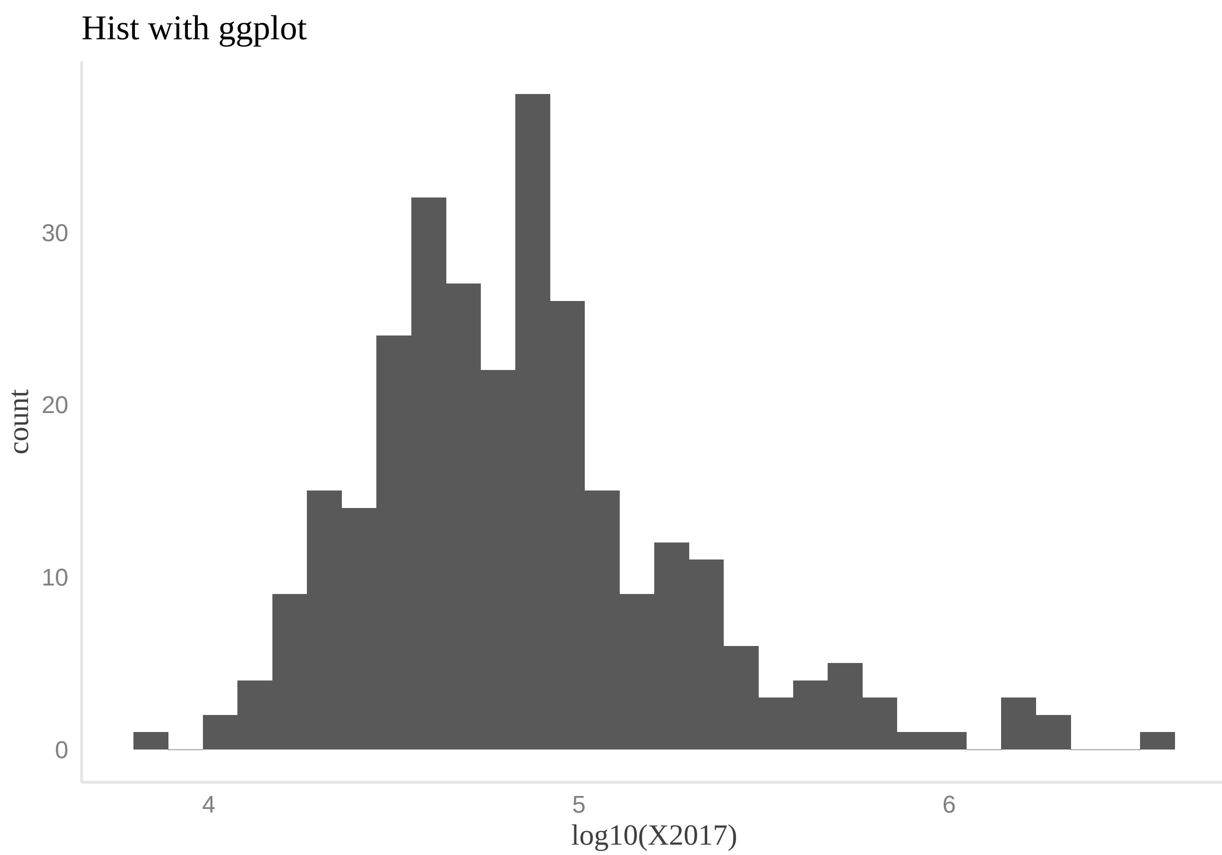
```r
ggplot(swedenCO2, aes(x = log10(X1990))) +
    geom_histogram() +
    ggtitle('Hist with ggplot')
```

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



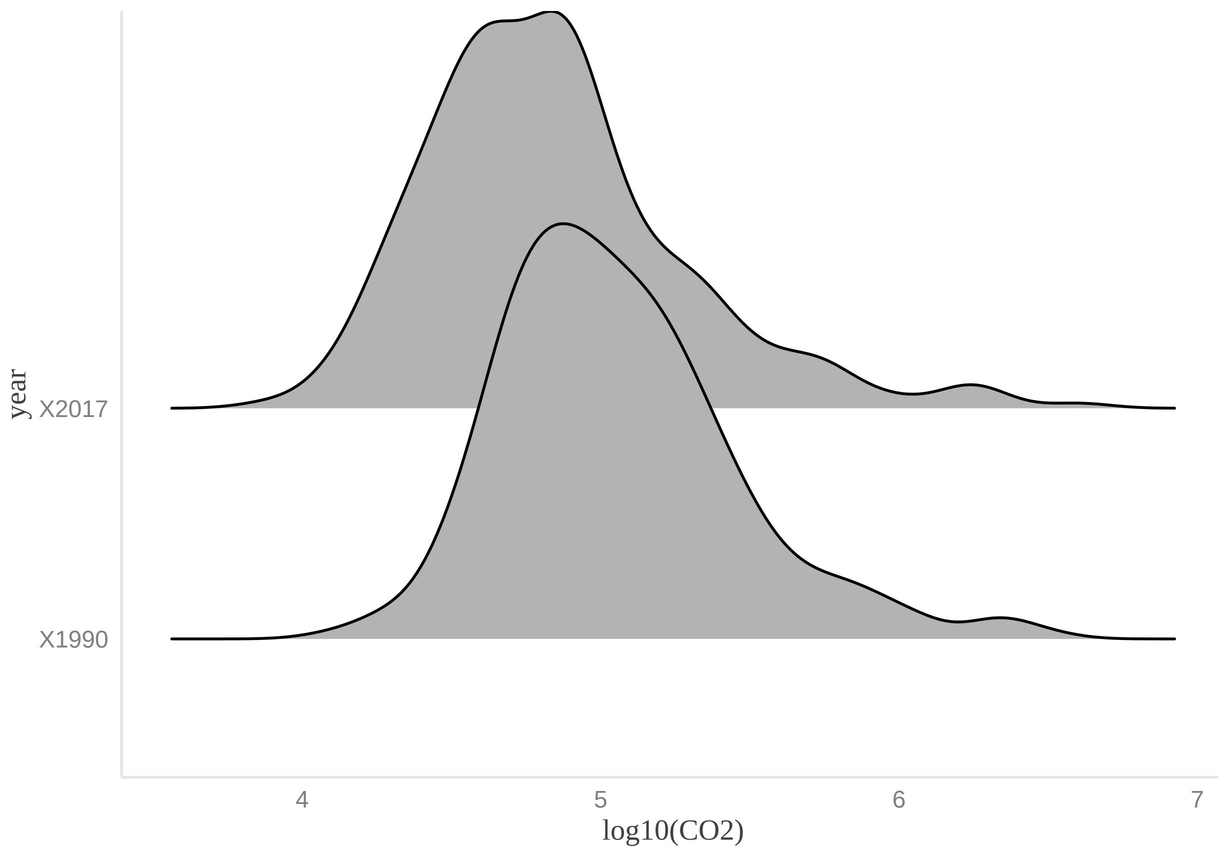

```r
ggplot(swedenCO2, aes(x = log10(X2017))) +
    geom_histogram() +
    ggtitle('Hist with ggplot')
```

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




```
swedenCO2 %>%
    select(X1990, X2017) %>%
    pivot_longer(cols = c(X1990, X2017), names_to = 'year', values_to = 'CO2') %>%
    ggplot(aes(x = log10(CO2), y = year)) +
    geom_density_ridges()
```

```
## Picking joint bandwidth of 0.107
```



The log-transformation did quite well. Let's see how the t-test does on the log-transformed data.

```
with(swedenCO2, (t.test(log10(X1990), log10(X2017), alternative='two.sided',
  conf.level=.95, paired=TRUE)))
```

```
##
##  Paired t-test
##
## data:  log10(X1990) and log10(X2017)
## t = 26.586, df = 289, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1958866 0.2272086
## sample estimates:
## mean of the differences
##               0.2115476
```

Now the p-value is as small as it was from the Wilcoxon Signed-rank test (and probably as small as R will return). The log-transformation enabled the use of the parametric test, although one cannot conclude from this example if the parametric test on transformed data is more powerful than non-parametric tests on the non-transformed data.

From previous studies, I've learned that the Wilcoxon Signed-rank tests should only be done on data that are symmetrical, and the raw data were clearly not symmetrical. Let's see if I can do a sign test instead on the raw data.

```r
activatePkgs('rstatix')
```

```
## Loading required package: rstatix
```

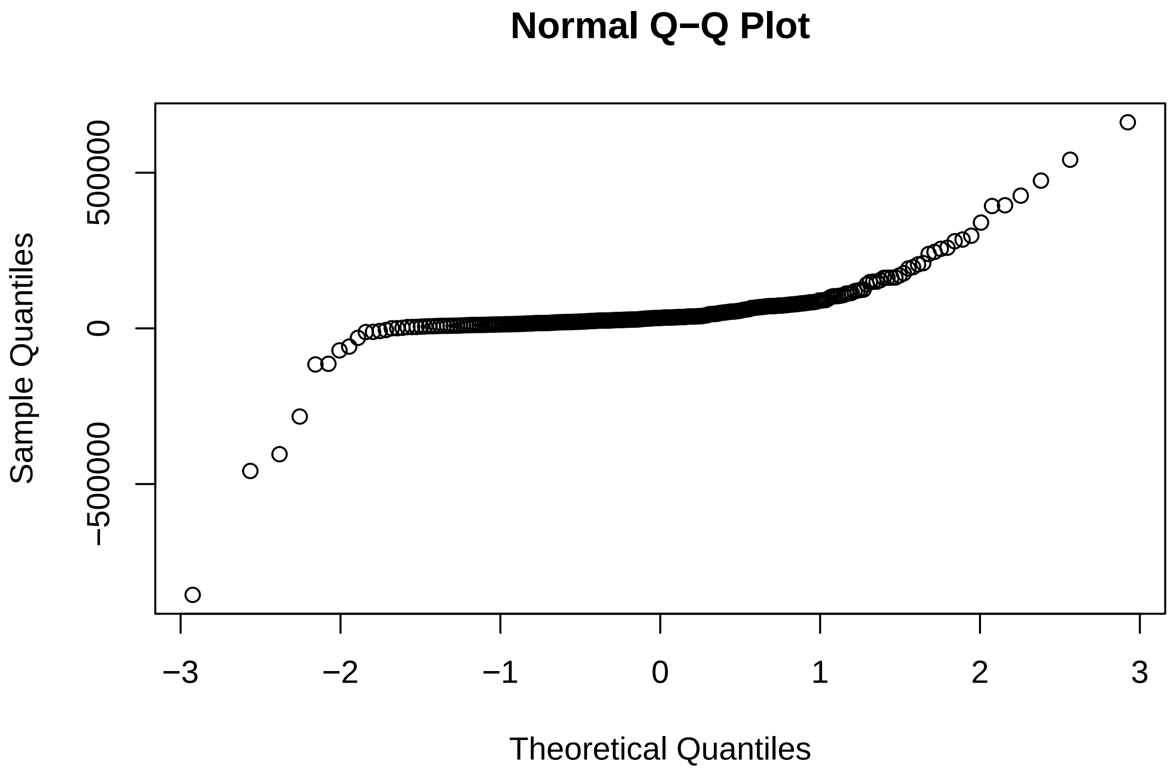
```r
swedenCO2 %>%
    select(X1990,X2017) %>%
    pivot_longer(cols = c(X1990, X2017), names_to = 'year', values_to = 'CO2') %>%
    pairwise_sign_test(CO2 ~ year)
```

```
## # A tibble: 1 x 10
##   .y.   group1 group2    n1    n2 statistic    df        p    p.adj p.adj.signif
## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 CO2   X1990  X2017    290   290       277   290 1.32e-65 1.32e-65 ****
```

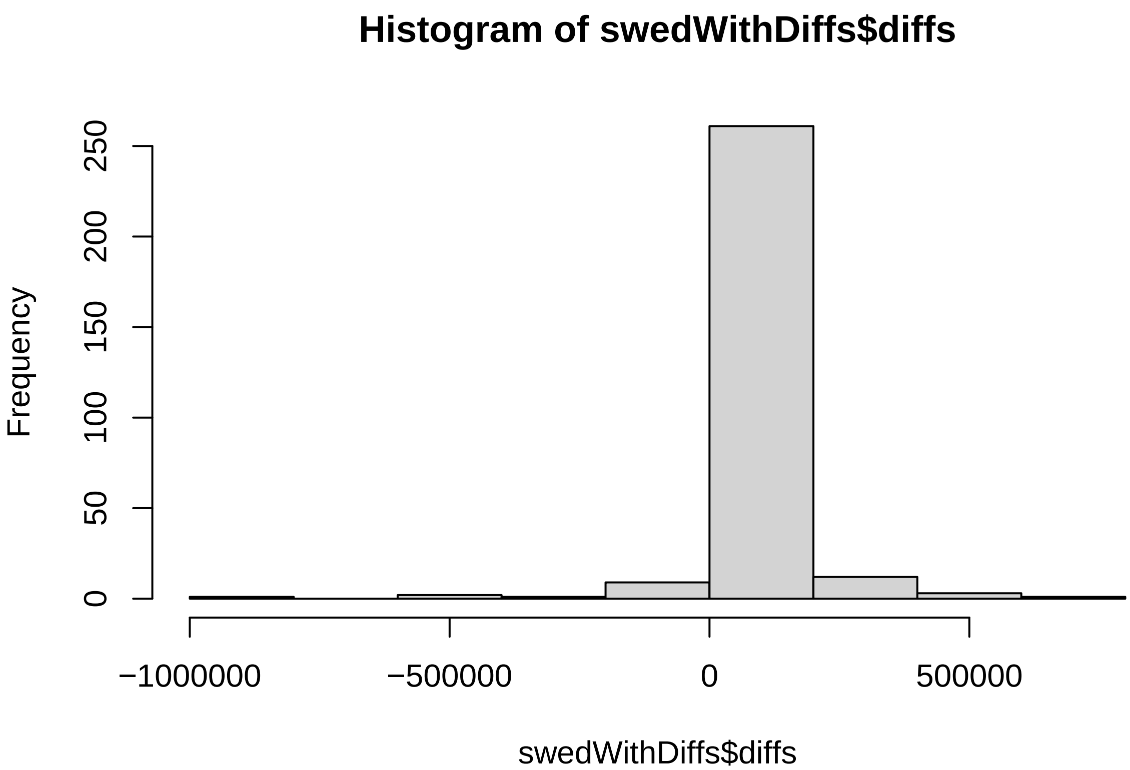
I did not find a `base` version of the sign test, but the `rstatix` package provided a version. This version, as best as I could tell, only accepts data in the formula syntax, so the data had to be stacked before it could be piped through the function. However, the results were a p-value that is many orders of magnitude lower than that given by the previous r functions. Either the previous r functions do not provide the same level of assurance, or the sign test is much more powerful given that the data do not satisfy the assumptions for using the Wilcoxon Signed-Rank test. In any case, the sign test did not appear to perform worse.

Upon reading more on the Wilcoxon Signed-Rank test, the assumption is that the differences between the two groups are symmetrical (reference here). Thus, let me take the differences and try to see if they are approximately symmetrical.

```r
swedWithDiffs <-
  swedenCO2 %>%
  mutate(diffs = na_remove(X1990 - X2017)) %>%
  select(diffs) # %>%
  # Unable to pipe directly into qqnorm, but have to save the object and call separately
  # qqnorm(.$diffs)

qqnorm(swedWithDiffs$diffs)
```



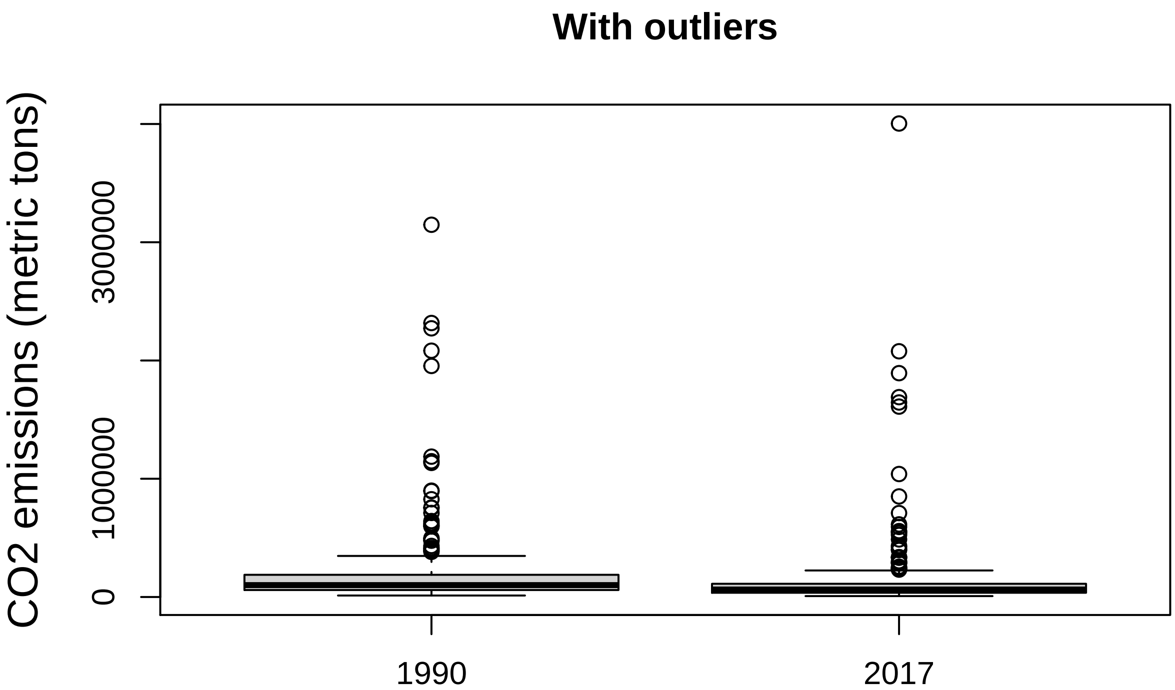

```
hist(swedWithDiffs$diffs)
```



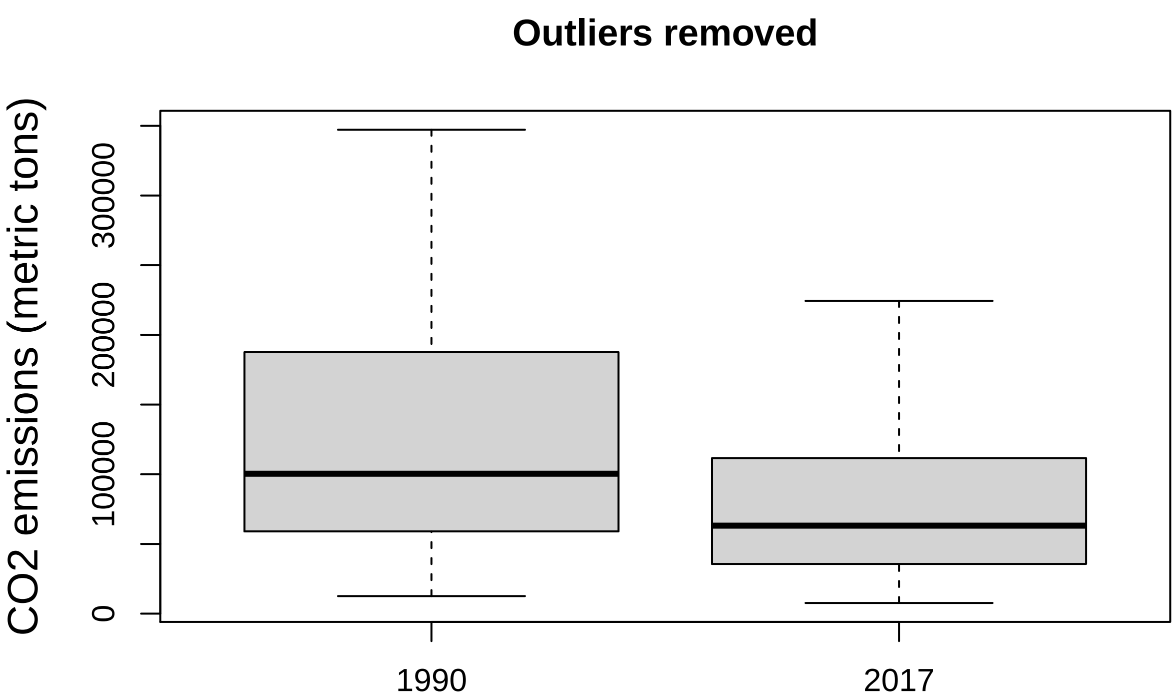
The histogram indicates the data are approximately symmetrical, but certainly not normally distributed (as also indicated by the qqplot).

Plotting the years side-by-side

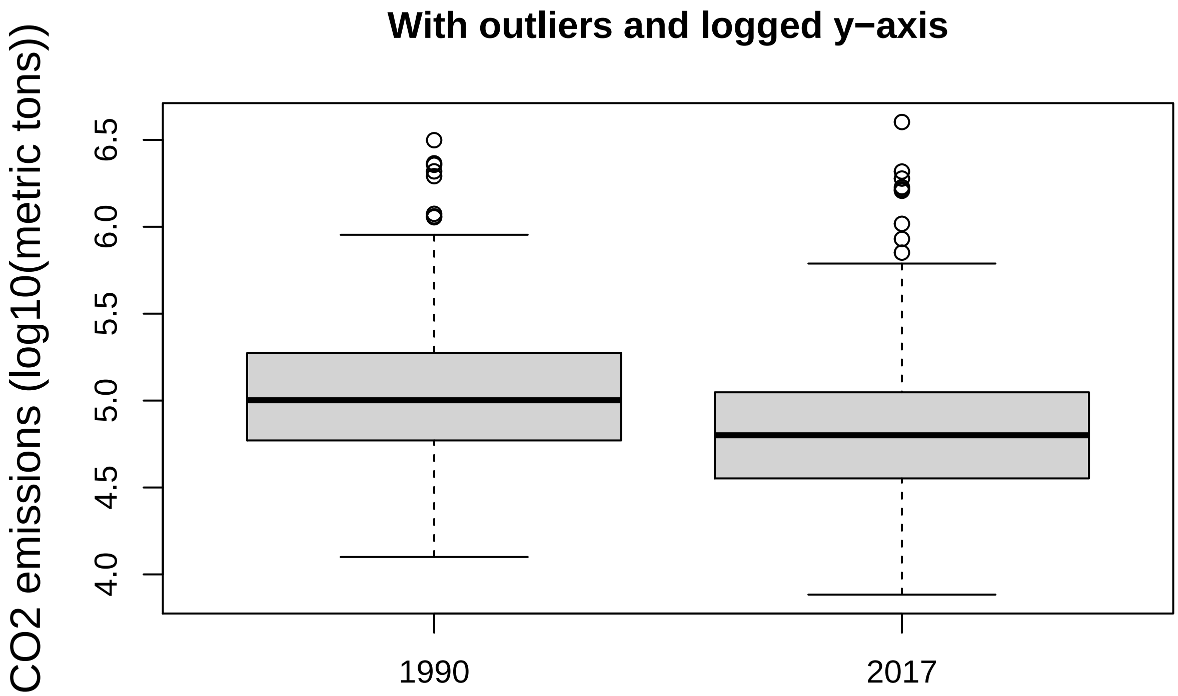
```
attach(swedenCO2)

boxplot(X1990, X2017, names=c("1990", "2017"), outline = TRUE, ylab="CO2 emissions (metric tons)", cex.l
```



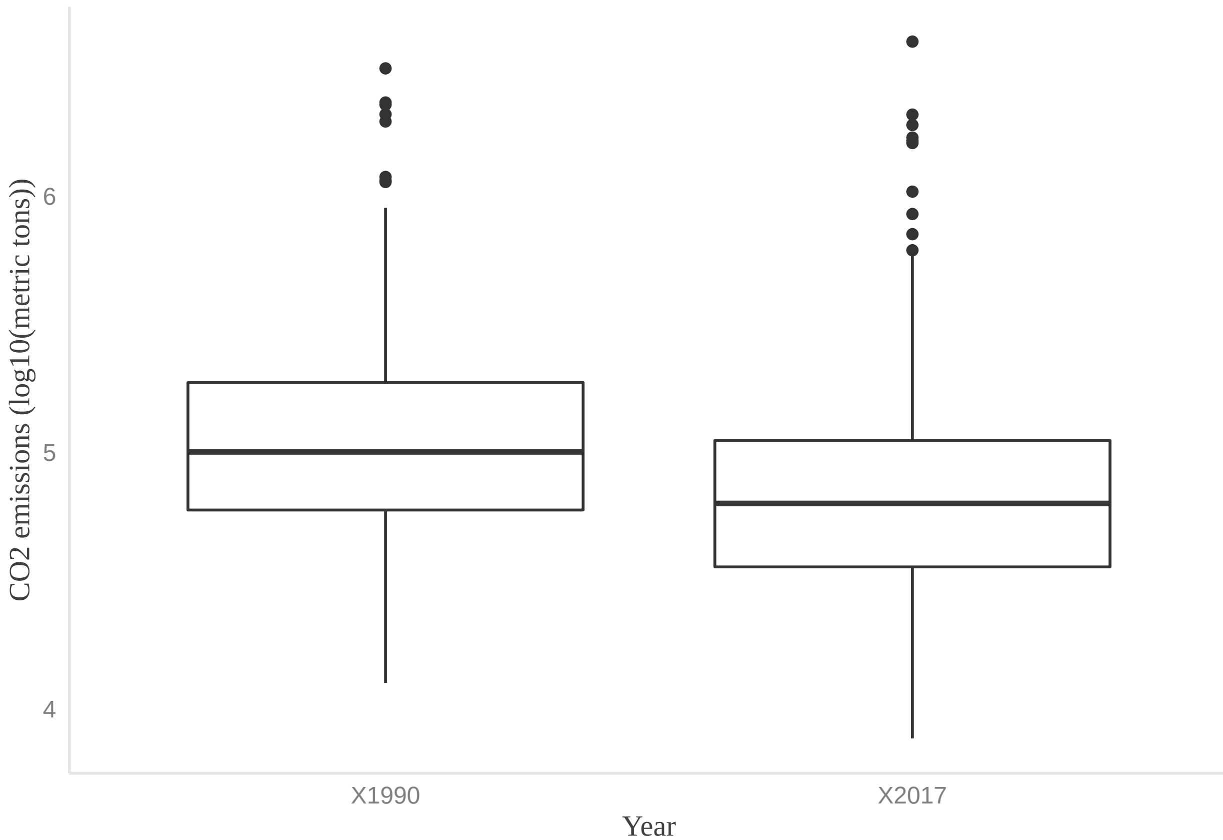

```
boxplot(X1990, X2017, names=c("1990", "2017"), outline = FALSE, ylab="CO2 emissions (metric tons)", cex
```



```r
boxplot(log10(X1990), log10(X2017), names=c("1990", "2017"), outline = TRUE, ylab="CO2 emissions (log10
```




```r
detach(swedenCO2)


swedenCO2 %>%
  select(X1990, X2017) %>%
  mutate(X1990 = log10(X1990), X2017 = log10(X2017)) %>%
  pivot_longer(cols = c(X1990, X2017), names_to = 'Year', values_to = 'CO2 emissions (log10(metric tons
  ggplot(aes(Year, `CO2 emissions (log10(metric tons))`)) +
  geom_boxplot()
```



Testing for differences in birth weight based on mothers' smoking clasification

I copied the data from the .html exercise file then manipulated in VS Code Insiders, then brought the manipulated data here.

```
bwBySmoke <- data.frame(
    `Birth weight` = c(
        3.18, 2.74, 2.9, 3.27, 3.65, 3.42, 3.23, 2.86, 3.6, 3.65, 3.69, 3.53, 2.38, 2.34, 3.99, 3.89, 3
        )
    , `Smoking habit` = c(
        'Heavy smokers', 'Heavy smokers', 'Heavy smokers', 'Heavy smokers', 'Heavy smokers', 'Heavy smo
        )
    )
```

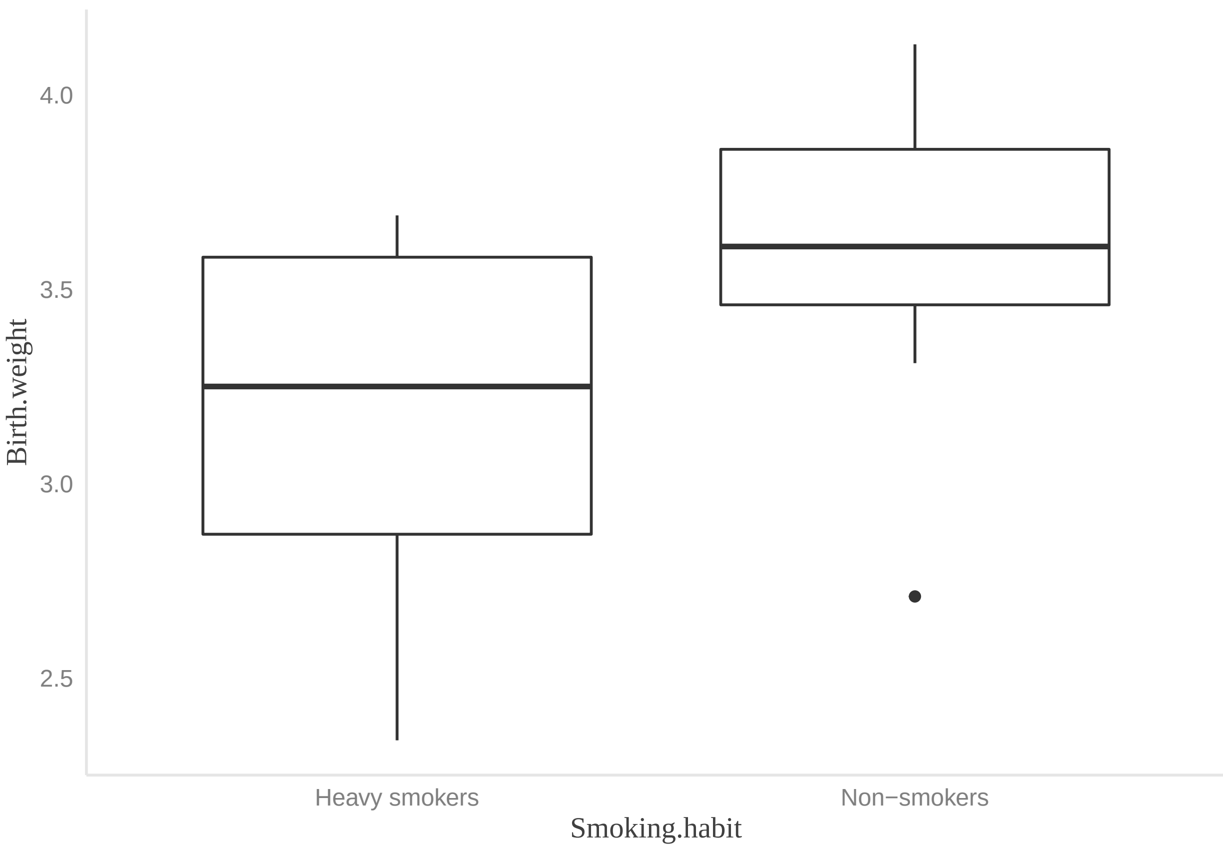
Exploratory Data Analysis

```
bwBySmoke %>%
  group_by(Smoking.habit) %>%
  summarize(
    n = n()
    , min = min(Birth.weight, na.rm = T)
    , q02 = quantile(Birth.weight, 0.02)
    , q16 = quantile(Birth.weight, 0.16)
    , median = median(Birth.weight, na.rm = T)
```

```
    , mean = mean(Birth.weight, na.rm = T)
    , q84 = quantile(Birth.weight, 0.84)
    , q98 = quantile(Birth.weight, 0.98)
    , max = max(Birth.weight, na.rm = T)
    , iqr = IQR(Birth.weight, na.rm = T)
    , mad = mad(Birth.weight, na.rm = T)
    , sd = sd(Birth.weight, na.rm = T)
    , sem = sd / sqrt(n)
    )
```

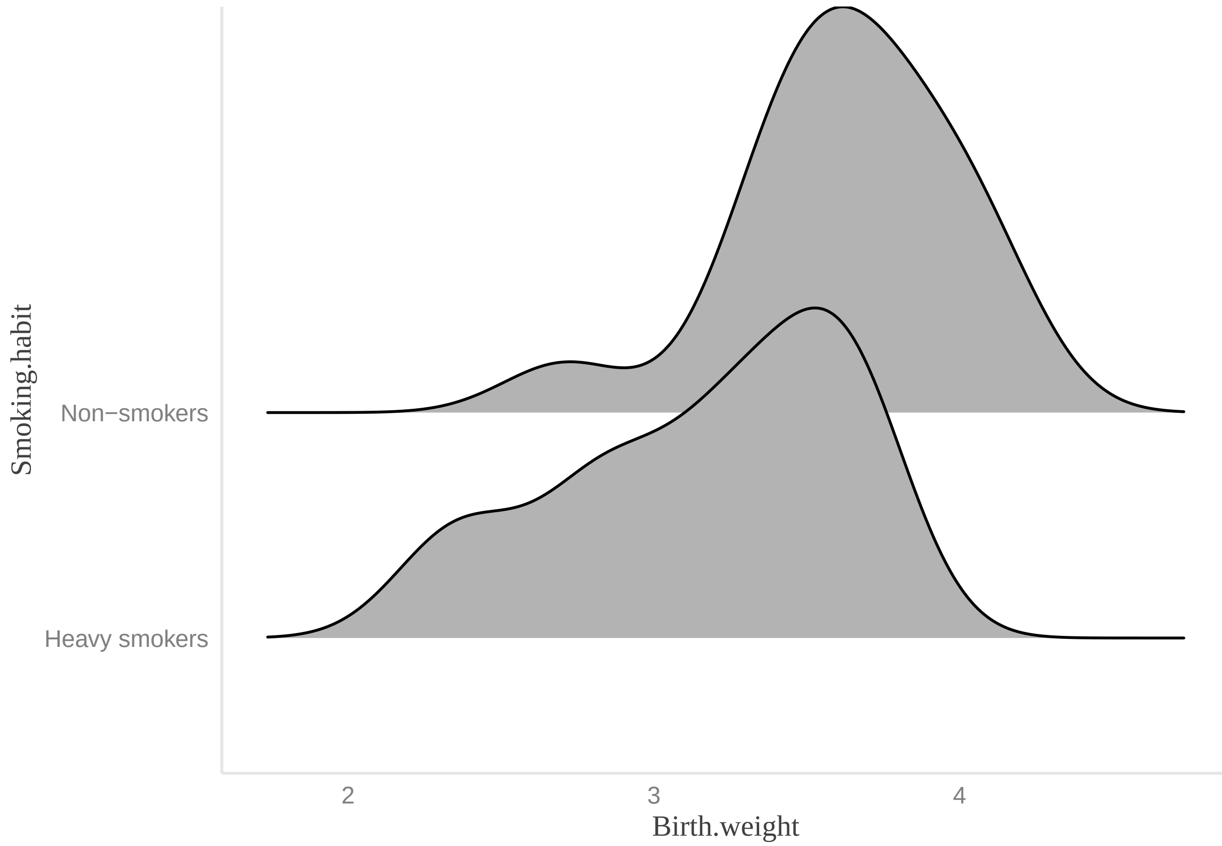
```
## # A tibble: 2 x 14
##   Smoking.habit     n   min   q02   q16 median  mean   q84   q98   max   iqr
##   <chr>         <int> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Heavy smokers    14  2.34  2.35  2.75   3.25  3.17  3.65  3.68  3.69 0.712
## 2 Non-smokers      15  2.71  2.88  3.37   3.61  3.63  3.97  4.12  4.13 0.400
## # ... with 3 more variables: mad <dbl>, sd <dbl>, sem <dbl>
```

```
bwBySmoke %>%
  ggplot(aes(x = Smoking.habit, y = Birth.weight)) +
  geom_boxplot()
```

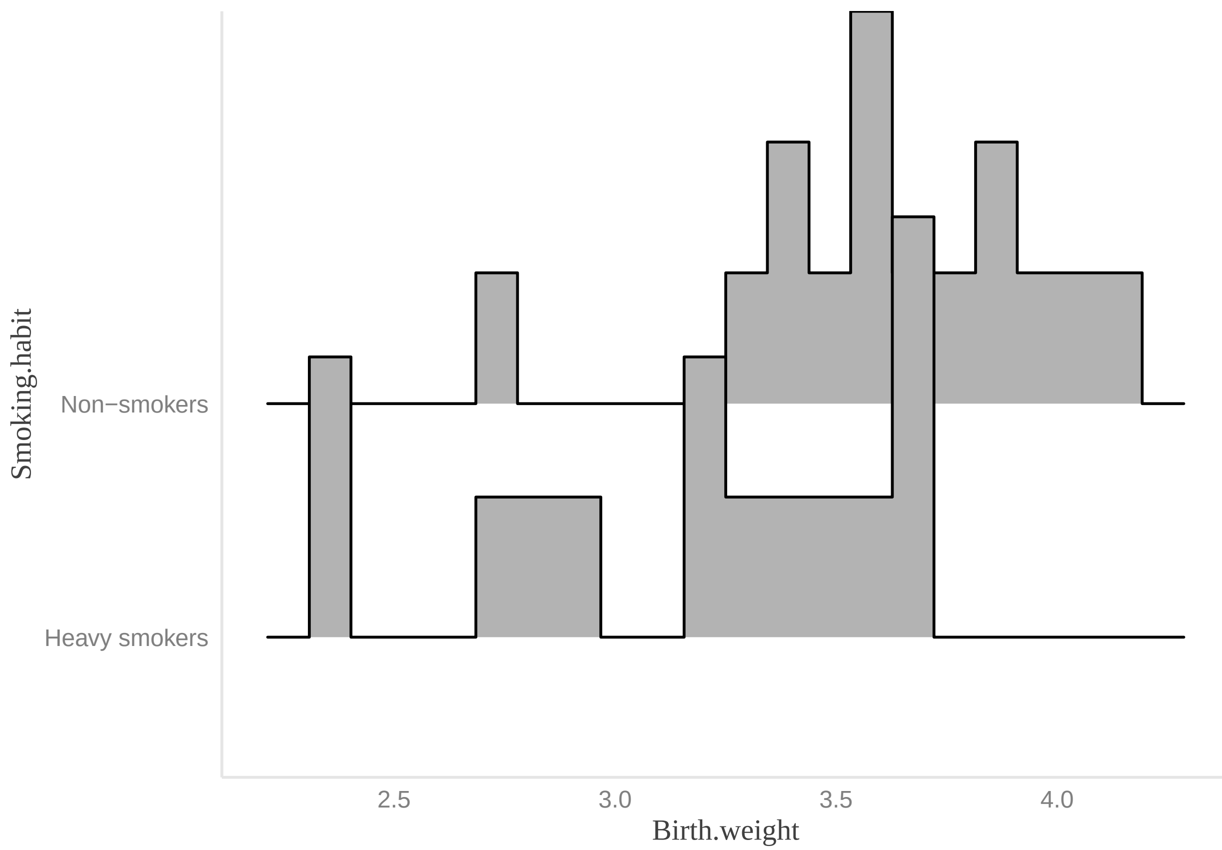



```
bwBySmoke %>%
  ggplot(aes(y = Smoking.habit, x = Birth.weight)) +
  geom_density_ridges()
```

```
## Picking joint bandwidth of 0.201
```




```
bwBySmoke %>%
  ggplot(aes(y = Smoking.habit, x = Birth.weight)) +
  geom_density_ridges(stat = 'binline', bins = 20)
```



These data do not appear to be normally distributed and thus I will not use the t-test. Rather I will use a non-parametric alternative. Specifically, I will assess if it is valid to use a a Wilcoxon Rank-Sum / Mann-Whitney U test.

According to this reference regarding the assumptions of a Wilcoxon Rank-Sum / Mann-Whitney U test, one can compare median ranks if the distributions have similar shapes, but only mean ranks otherwise.

Since I am not certain if the distributions are similar enough to compare median ranks, I will do a Wilcoxon Rank-Sum / Mann-Whitney U test by comparing mean ranks.

Unfortunately, I did not find a way to compare mean ranks with this test and have posted to CrossValidated to see if someone can help. Instead, I will simply use the default behavior, which I believe compares median ranks.

```
with(bwBySmoke, wilcox.test(Birth.weight~Smoking.habit))
```

```
## Warning in wilcox.test.default(x = c(3.18, 2.74, 2.9, 3.27, 3.65, 3.42, : cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Birth.weight by Smoking.habit
## W = 45.5, p-value = 0.01001
## alternative hypothesis: true location shift is not equal to 0
```

The test statistic for the Wilcoxon Rank-Sum test, **W is 30.5**. The **p-value is 0.001238**. The **group sizes are 14** for the 'Heavy-smokers' group **and 15** for the 'Non-smokers' group. The **median birth weights are 3.25 and 3.61 kg** for the 'Heavy-smokers' and 'Non-smokers' groups, respectively.

The conclusion is that heavy smokers give birth to lower-weight babies.

Key learnings

- Scoping is different when knitting markdown files vs developing them directly in RStudio
- To knit to pdf, one must have some distribution of LaTeX. This may require, for example, installing MiKTex. After installation, one should check for and install all updates.
- To silence a warning while knitting to pdf that has to do with how plots are cropped, one may need to install Ghostscript and add its executable file to the Windows PATH variable.
- If warnings/errors are still arising after installing MiKTeX and/or Ghostscript, it may be worth re-starting the computer and/or removing those installations, re-installing while all other applications are shut down, and then rebooting the computer.
- The `attach` function is probably what is behind the scenes for the pipe function in the tidyverse… It makes a dataset available for reference in subsequent calls without needing to use subsetting operators (`$` and `[]` and `[[]]`). This is probably also what is behind the scenes when activating a dataset in the `Rcmdr` UI. The `detach` function removes it.

Unresolved questions

- Isn't an assumption for the Wilcoxon signed-rank test that the distribution of paired differences are symmetrical? If not symmetrical, then one should do a Sign test instead, right?
- Is the Wilcoxon rank-sum test the same as the Mann-Whitney U test?
- With the Wilcoxon rank-sum test / Mann-Whitney U test, don't you need to examine if the distributions of the compared groups are similar? If they are similar then you can compare medians, but if they are not similar you compare means, right?
- Why is the df lower than 8 when doing the t test on width~sex but 8 (as expected) when doing length~sex?
- If doing a paired t-test but homogeneity of variances cannot be assumed, is the alternative test also a Welch t-test, or is a Welch t-test only for unpaired data that violates the assumption of homogeneity of variances?
- When doing a Wilcoxon Signed-Rank test using `Rcmdr`, one has the options to choose the 'default', 'exact', 'normal approximation', or 'normal approximation with continuity correction' test. When would one select the different options?
- Why can't I pipe directly into the `qqnorm` function with a selected result column?
- How does one formally/objectively assess if the distributions of differences are symmetrical, as is an assumption for the Wilcoxon Signed-Rank test?
- How does one formally/objectively assess if the distributions are similarly-shaped, as is an assumption for the Wilcoxon Rank-Sum / Mann-Whitney U test?