HW6 resampling

AUTHOR
Darrell Sonntag

PUBLISHED
March 4, 2024

Read in the data we used for HW5

We will evaluate the vehicle emissions data we evaluated in HW5.

Let’s evaluate the ‘summer.2022.gas.NO’ dataset.

Let’s just look at the MY 1994 and newer vehicles

```r
summer.2022.gas <- read_csv("../../CE594R_25_data_science_class/data/summer.2022.gas.csv")
```

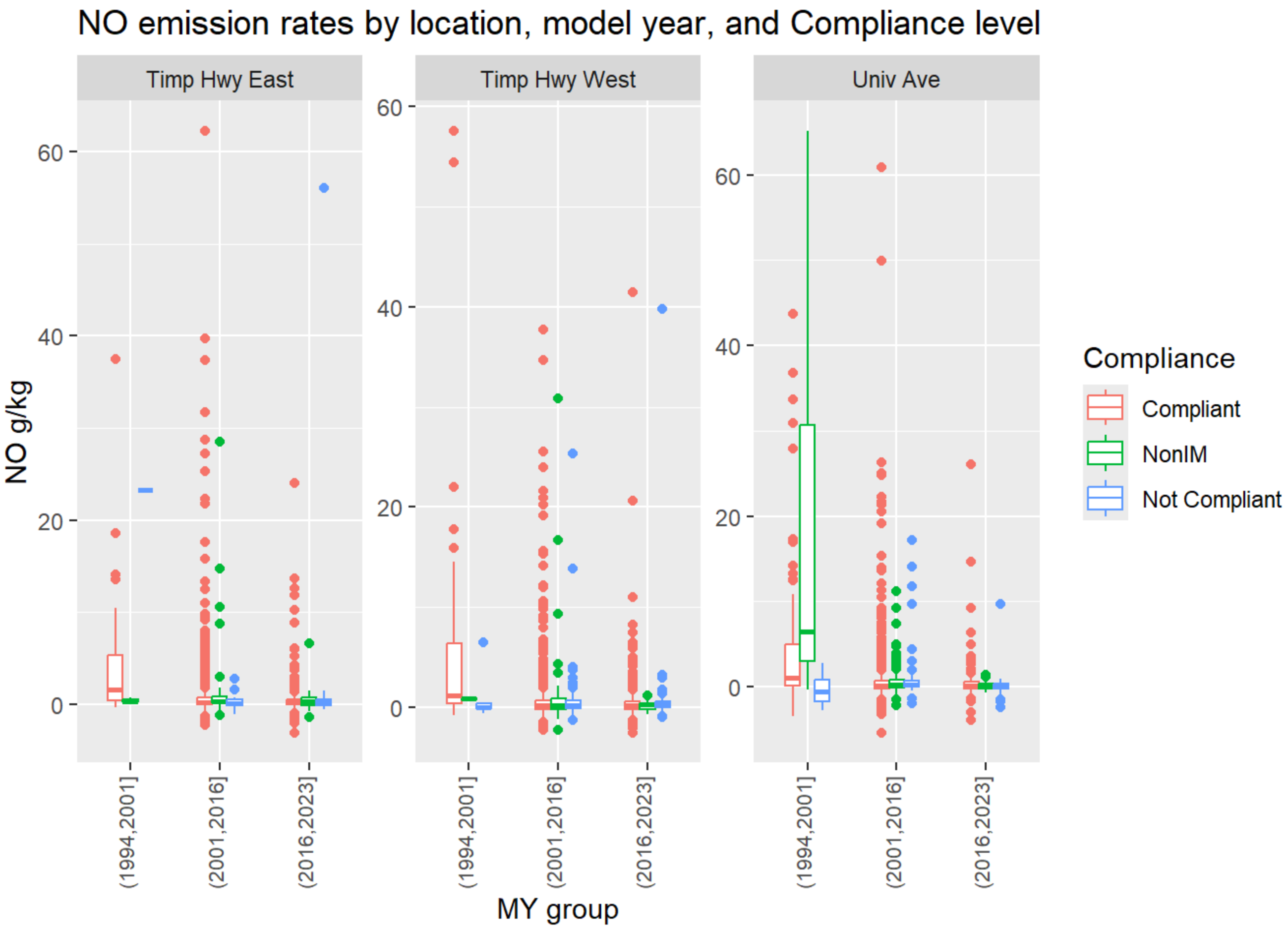
```
Rows: 32256 Columns: 36
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (22): LICENSE, DATE, VIN, Vehicle Type, Make, Model, Fuel, FuelGroup, W...
dbl   (9): Year, Zip, Registered Weight, SPEED, ACCEL, id, ER, Max GVWR, Cur...
lgl   (2): LICENSE_outofstate, Veh.info.corrected
date  (2): Last Emissions Date, Last Test Date Required
time  (1): TIME

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
summer.2022.gas.NO <- summer.2022.gas %>%
    filter(Compliance %in% c("Compliant", "Not Compliant","NonIM")) %>%
    filter(pollutant == 'NO') %>%
    filter(Year>1994) %>%
    filter(!is.na(ER))
```

Plot box plots of the data. With panels for each location, the model year on the x-axis, and different colors for each Compliance level (dodged). You can use the code below

```r
ggplot(data = summer.2022.gas.NO, aes(x = year_cuts, y = ER, color= Compliance)) +
geom_boxplot(position = position_dodge(width = 0.5),width=0.5) +
facet_wrap(.~ location, scales = "free_y") +
labs(title = "NO emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NO g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2))
```



```
ggsave("../../CE594R_data_science_R/figs/NO_boxplots.png")
```

```
Saving 7 x 5 in image
```

Calculate the means and 95% CI of the mean for each groups

Let’s calculate the means by 3 model year groups (using year_cuts), compliance status, and location

Then, let’s calculate the 95% CI using the t-distribution

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

```
NO.loc.comp.summary <- summer.2022.gas.NO %>%
  group_by(pollutant, location, Compliance,year_cuts) %>%
  summarize(mean = mean(ER, na.rm=T),median = median(ER,na.rm=T),
            sd=sd(ER,na.rm=T),n=sum(!is.na(ER)),
            min=min(ER,na.rm=T),max=max(ER,na.rm=T)) %>%
  mutate(tcrit = qt(.975,df=(n-1))) %>%
  mutate(bound = tcrit*sd/sqrt(n)) %>%
  mutate(lower.95 = mean-bound) %>%
  mutate(upper.95 = mean+bound)
```

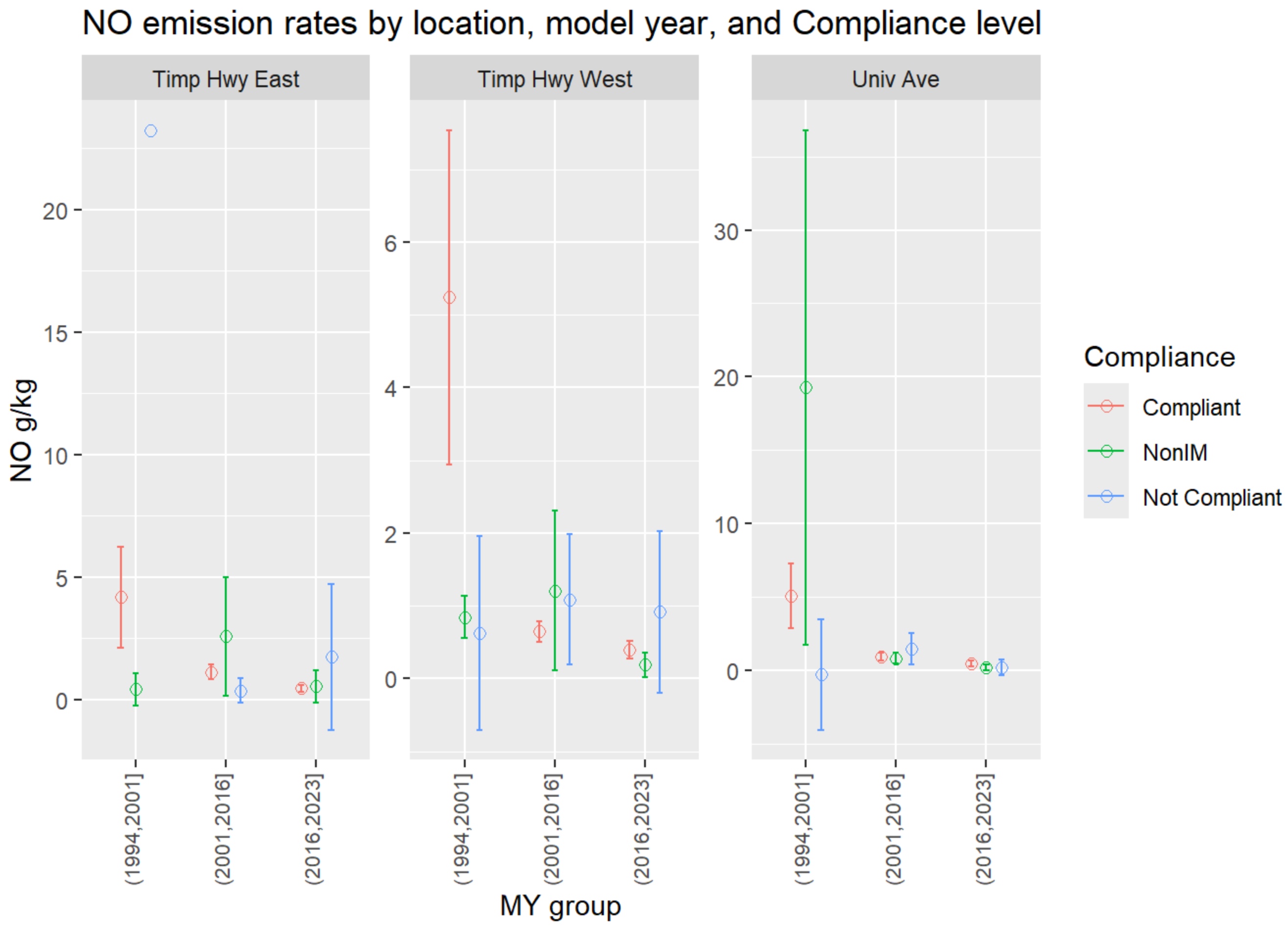
```
`summarise()` has grouped output by 'pollutant', 'location', 'Compliance'. You
can override using the `.groups` argument.

Warning: There was 1 warning in `mutate()`.
ℹ In argument: `tcrit = qt(0.975, df = (n - 1))`.
ℹ In group 3: `pollutant = "NO"`, `location = "Timp Hwy East"`, `Compliance =
  "Not Compliant"`.
Caused by warning in `qt()`:
! NaNs produced
```

Plot the means for each of the group with 95% CI

Plot Model year on the x-axis ER on the y-axis Separate panels for each location Dodge by Compliance

```
ggplot(data = NO.loc.comp.summary, aes(x = year_cuts, y = mean, color= Compliance)) +
geom_point(position = position_dodge(width = 0.5),size =2, shape=1) +
geom_errorbar(position = position_dodge(width = 0.5), aes(ymin = lower.95, ymax = upper.95), widt
facet_wrap(.~ location, scales = "free_y") +
labs(title = "NO emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NO g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2))
```



Question: Which error bars are the widest? Question: What is the n for these groups?

Answer:

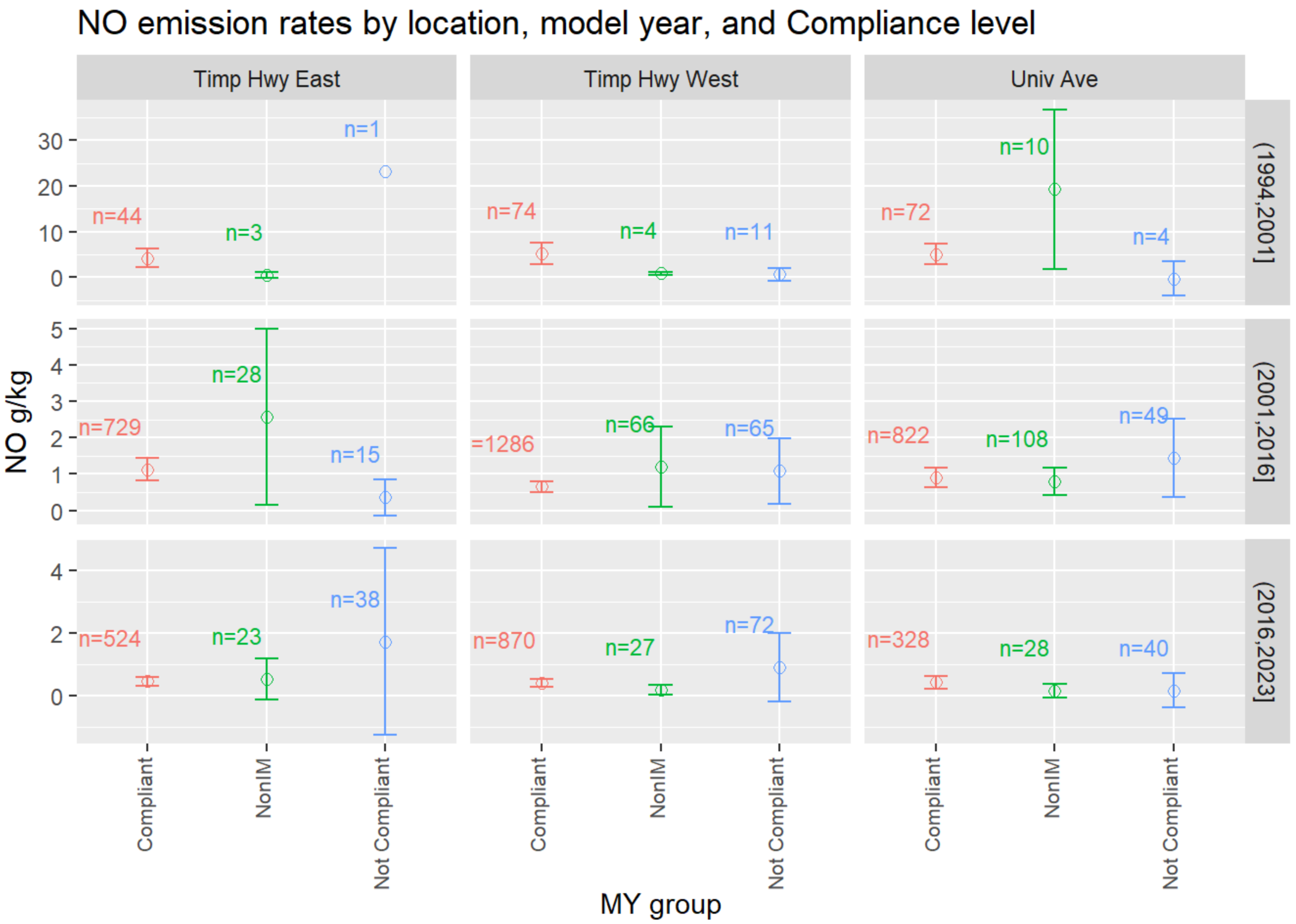
1994-2001 Compliant Timp Hwy West - 74 vehicles 1994-2001 NonIM - 10 vehicles

The wide confidence intervals for some of the groups, makes is difficult to observe differences in others.

Let’s plot the means again for each of the group with 95% CI

Except this time, let’s have separate panels for location AND model year group Plot Compliance on the x-axis Add the number of observations in each group using geom_text

```
ggplot(data = NO.loc.comp.summary, aes(x = Compliance, y = mean, color= Compliance)) +
geom_point(size =2, shape=1) +
geom_errorbar(aes(ymin = lower.95, ymax = upper.95), width = 0.2) +
facet_grid(year_cuts~ location, scales = "free_y") +
geom_text(aes(label=paste("n=",n,sep='')), size = 3,hjust = 1.1,vjust=-2)+
labs(title = "NO emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NO g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2),legend.position = "none"
```



```
ggsave("../../CE594R_data_science_R/figs/NO_t_conf.png")
```

```
Saving 7 x 5 in image
```

Question: How many of the confidence intervals include 0?

Answer:

9 of the 27 groups

Notice the 95% confidence intervals using the t-distribution are very large—

for some groups, the large N may compensate for the data being highly skewed and non-normal.

However, for the samples < 100, the distribution of the means may not be well approximated with a t-distribution.

Let’s use resampling to calculate 95% confidence intervals that are useful for non-normal and small sample size data

First let’s use the loop method.

We have 3 locations X 3 compliance levels X 3 model year groups = 27 unique groups

We could subdivide the data into 27 unique groups…and then resample from each (like we did in class)

And then resample using a for loop

However, that would be difficult with 27 unique groups.

First, I createed a function that does the resampling

```
resample_cars <- function(data.i){

                  resample <- data.frame()
                  for (i in 1:100){
                    resample.i <- data.i %>%
                        slice_sample(n=nrow(data.i),replace=TRUE) %>%
                        mutate(sample = i)

                    resample <- bind_rows(resample,resample.i)
                    }
                  return(resample)
                  }
```

When I tried the above function it took way too long…

So, I used the rep_sample_n from the moderndive package https://moderndive.com/8-confidence-intervals.html#original-workflow

```
resample_cars <- function(data.i){

                  resample <- rep_sample_n(data.i,size = nrow(data.i),reps=1000,replace = TRUE )
                  return(resample)
                  }
```

Apply the function use the group_by, group_split, map, and list_rbind

Info on the group_split here. https://dplyr.tidyverse.org/reference/group_split.html

map

https://purrr.tidyverse.org/reference/map.html

Map() I applied the function I created above. Map returns a list of output. I then used list_rbind() to bind all the list elements of the back together into a dataframe/tibble).

You could also use the base R version of split(), and then use lapply()

```
resample.NO <- summer.2022.gas.NO  %>%
                       group_by(location,year_cuts,Compliance) %>%
                       group_split() %>%
                       map(resample_cars) %>%
                       list_rbind()
```

Summarize the means by replicate

```
names(resample.NO)
```

```
 [1] "replicate"               "LICENSE"
 [3] "DATE"                    "TIME"
 [5] "LICENSE_outofstate"      "Veh.info.corrected"
 [7] "VIN"                     "Vehicle Type"
 [9] "Make"                    "Model"
[11] "Year"                    "Fuel"
[13] "FuelGroup"               "Zip"
[15] "Weight Rating"           "Registered Weight"
[17] "Last Emissions Date"     "SPEED_FLAG"
[19] "SPEED"                   "ACCEL"
[21] "TAG_NAME"                "location"
[23] "id"                      "pollutant"
[25] "ER"                      "CO2_FLAG"
[27] "POLLUTANT_FLAG"          "County"
[29] "City"                    "Max GVWR"
[31] "GVWR Requirement"        "Current Year"
[33] "Required"                "Frequency"
```

```
[35] "Last Test Date Required" "Compliance"
[37] "year_cuts"
```

```
resample.NO.means <- resample.NO %>%
                      group_by(replicate,location,year_cuts,Compliance) %>%
                      summarize(sample_mean = mean(ER))
```

```
`summarise()` has grouped output by 'replicate', 'location', 'year_cuts'. You
can override using the `.groups` argument.
```

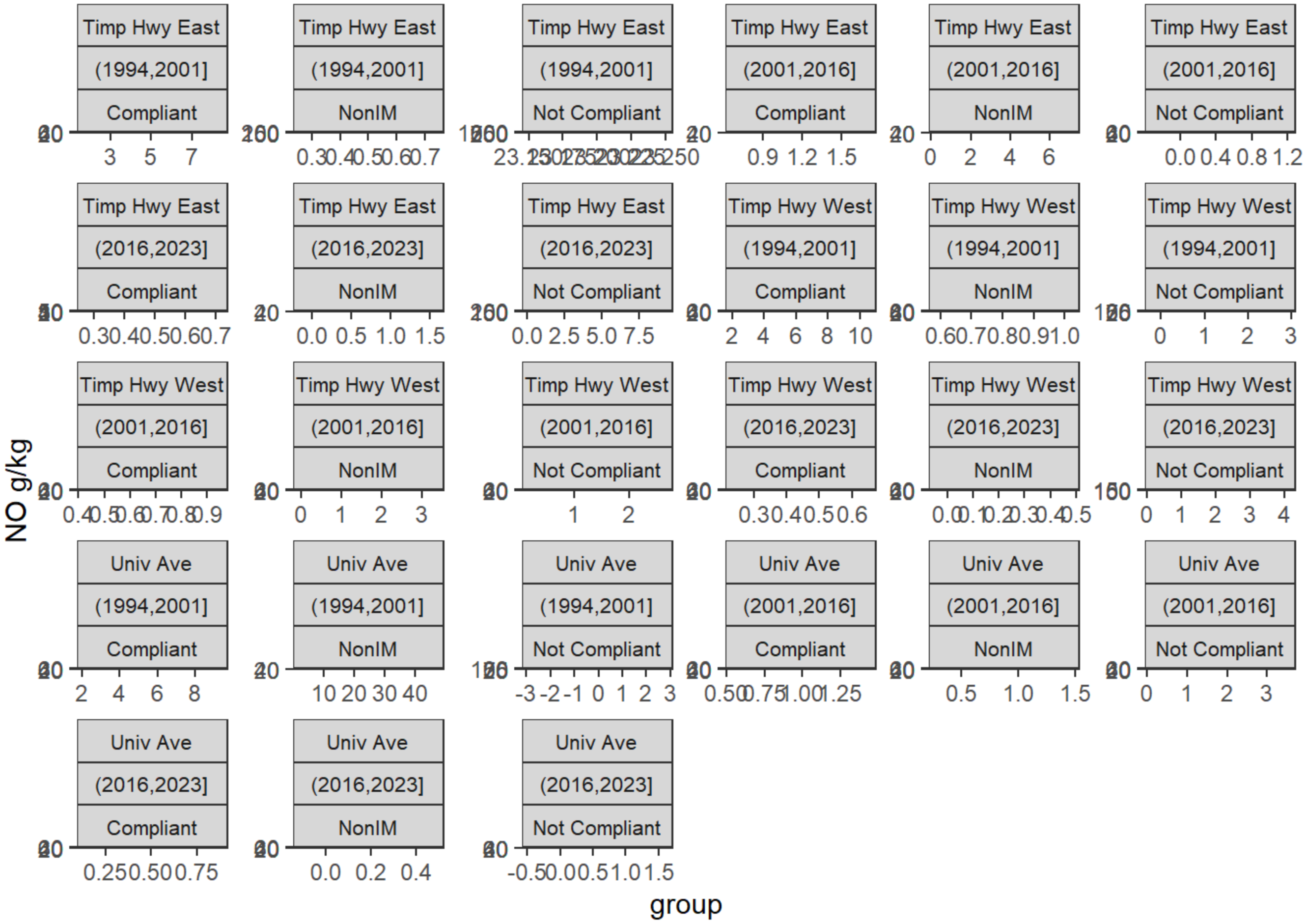
Calculate the 2.5 and 97.5% quantiles

```
resample.NO.intervals <- resample.NO.means %>%
                          group_by(location,year_cuts,Compliance) %>%
                          summarize(mean = mean(sample_mean),
                            lower.95 = quantile(sample_mean,0.025),
                            upper.95 = quantile(sample_mean,0.975))
```

```
`summarise()` has grouped output by 'location', 'year_cuts'. You can override
using the `.groups` argument.
```

Plot histograms of the resample means for each group, with the 95% CIs like the figure below

```
ggplot(data = resample.NO.means,aes(x = sample_mean,fill=Compliance )) +
  geom_histogram(bins=50)+
  geom_vline(data=resample.NO.intervals,aes(xintercept = lower.95)) +
  geom_vline(data=resample.NO.intervals,aes(xintercept = upper.95)) +
  facet_wrap(.~location + year_cuts+Compliance, scales='free') +
  theme_bw()+
  labs(x='group', y= 'NO g/kg')  +
  theme(strip.text = element_text(size=8),legend.position="none")
```



```
ggsave("../../CE594R_data_science_R/figs/NO_resampling_means.png")
```

```
Saving 7 x 5 in image
```

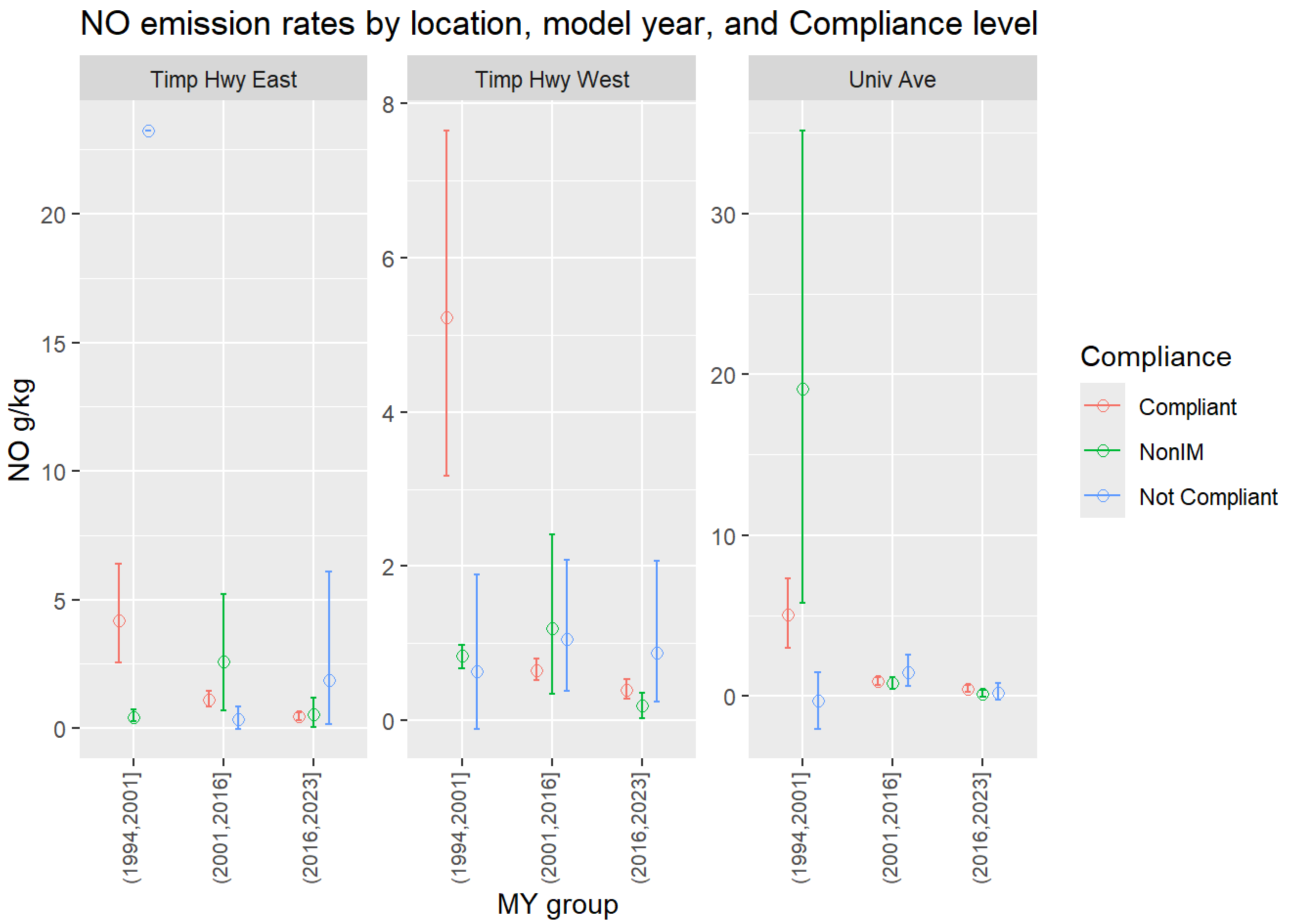
Question: Why do some of the resampling mean distributions have multiple modes, or clusters?

Hint: Look at the box plot figure in

Answer: Some of the samples have 1-2 very large outliers. For example: Timp Hwy East - Not compliant 2016-2023, has 38 vehicles, with one very large outlier. If the resample includes multiple occurances of the high emitting vehicle, the mean is significantly higher. So, each cluster represents the high emitting vehicle being present 1, 2, 3, 4 or 5 times in the resample.

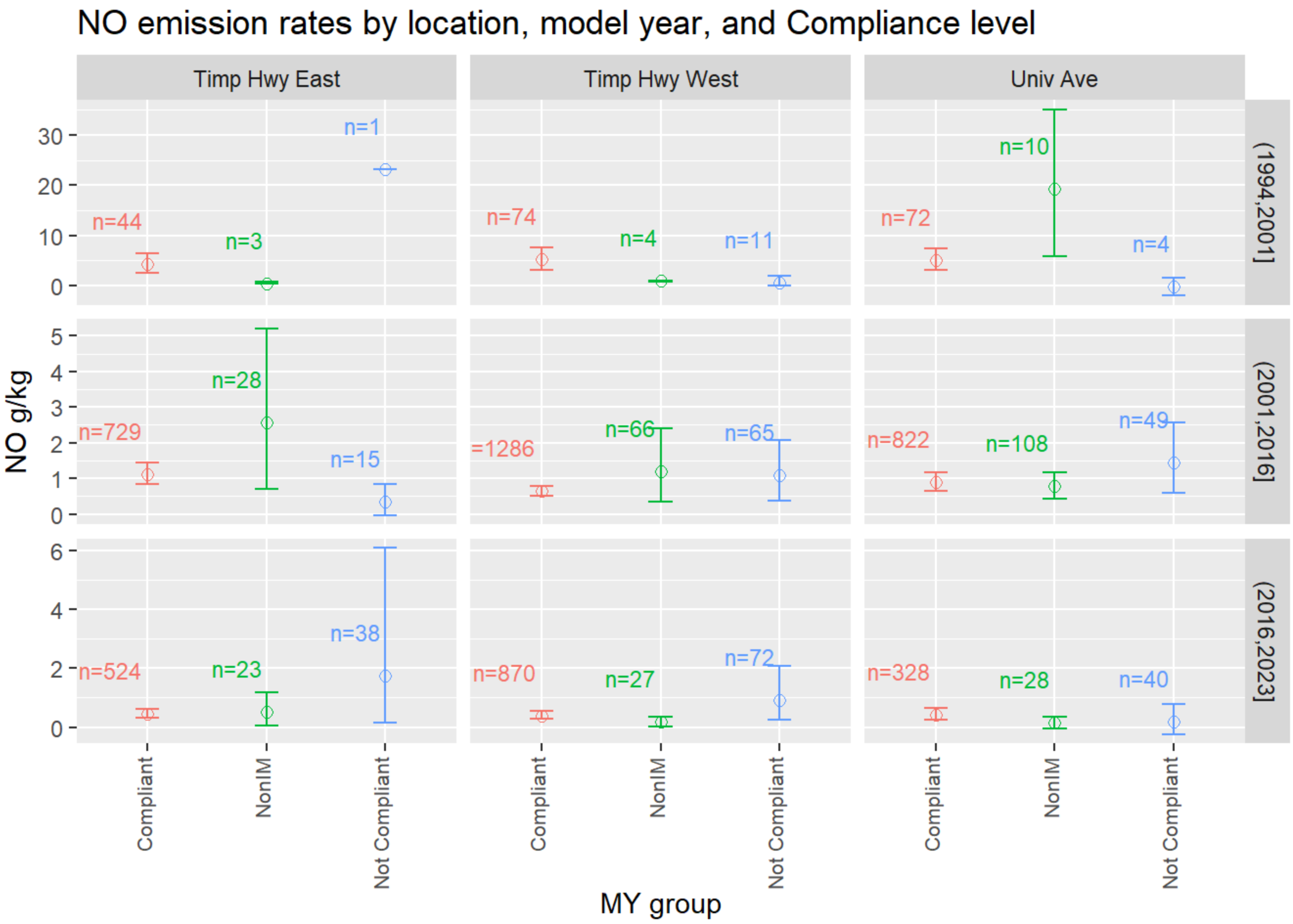
Plot the emission rates with error bars for the quantiles

```
ggplot(data = resample.NO.intervals, aes(x = year_cuts, y = mean, color= Compliance)) +
geom_point(position = position_dodge(width = 0.5),size =2, shape=1) +
geom_errorbar(position = position_dodge(width = 0.5), aes(ymin = lower.95, ymax = upper.95), widtl
facet_wrap(.~ location, scales = "free_y") +
labs(title = "NO emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NO g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2))
```



Plot it using the panels

```
ggplot(data = NO.loc.comp.summary,
       aes(x = Compliance, y = mean, color= Compliance)) +
geom_point(size =2, shape=1) +
geom_errorbar(data = resample.NO.intervals,
              aes(ymin = lower.95, ymax = upper.95), width = 0.2) +
facet_grid(year_cuts~ location, scales = "free_y") +
geom_text(aes(label=paste("n=",n,sep='')), size = 3,hjust = 1.1,vjust=-2)+
labs(title = "NO emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NO g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2),legend.position = "none"
```



```
ggsave("../../CE594R_data_science_R/figs/NO_resampling_conf.png")
```

```
Saving 7 x 5 in image
```

How many of groups have confidence intervals that include zero?

Now just 5 of the 27 groups have confidence intervals that include zero (dropped from 9).

Graph them side by side

```
names(NO.loc.comp.summary)
```

```
 [1] "pollutant"  "location"   "Compliance" "year_cuts"  "mean"
 [6] "median"     "sd"         "n"          "min"        "max"
[11] "tcrit"      "bound"      "lower.95"   "upper.95"
```

```
names(resample.NO.intervals)
```

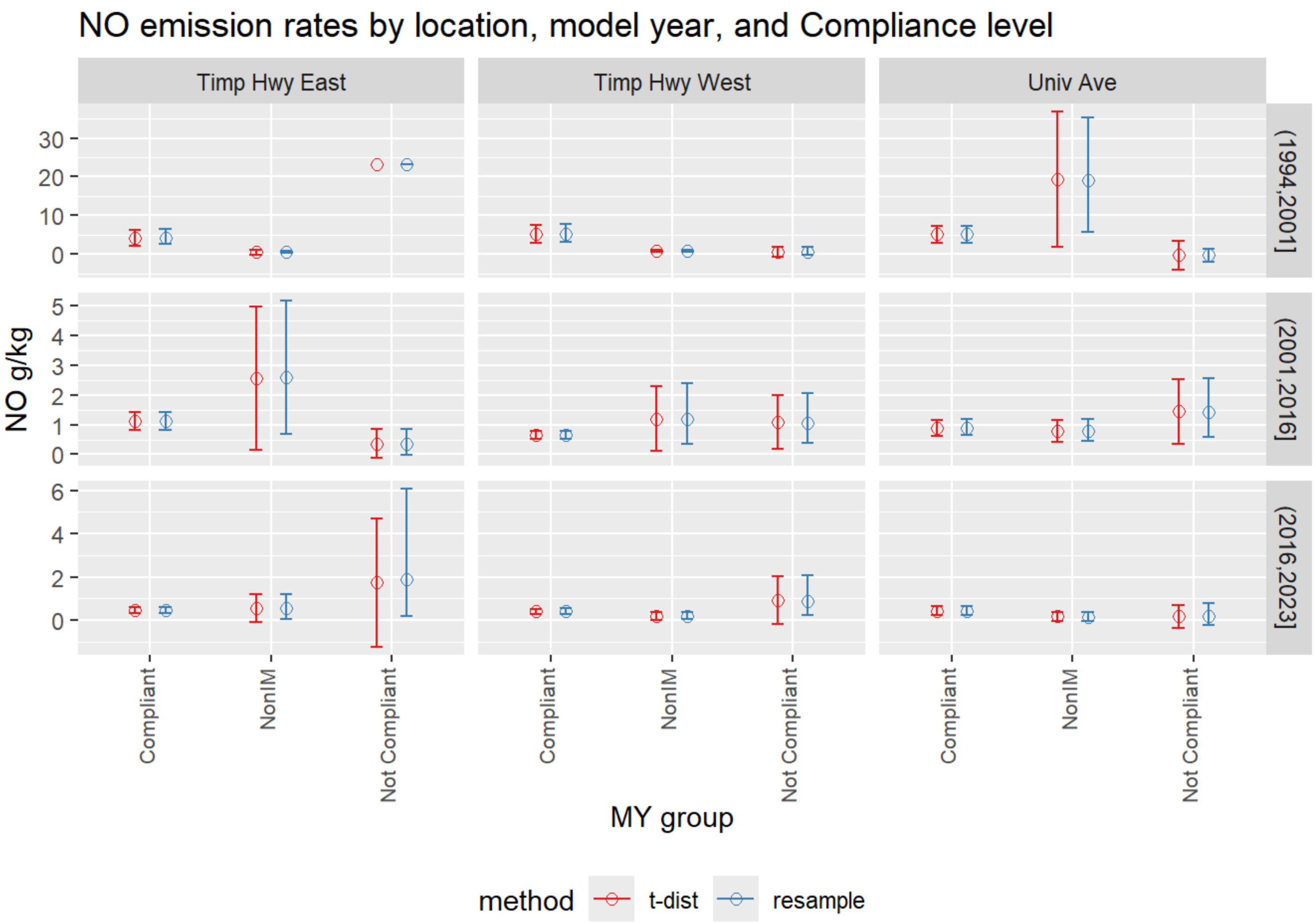
```
[1] "location"   "year_cuts"  "Compliance" "mean"       "lower.95"
[6] "upper.95"
```

```
NO.loc.comp.summary$method = 't-dist'

resample.NO.intervals$method = 'resample'

compare.conf.intervals <- bind_rows(NO.loc.comp.summary,resample.NO.intervals) %>%
                          mutate(method = factor(method,levels=c('t-dist','resample'),ordered=TRU


ggplot(data = compare.conf.intervals,
       aes(x = Compliance, y = mean, color= method)) +
geom_point(size =2, shape=1, position = position_dodge(width = 0.5)) +
geom_errorbar(aes(ymin = lower.95, ymax = upper.95), width = 0.2,position = position_dodge(width =
facet_grid(year_cuts~ location, scales = "free_y") +
scale_color_brewer(palette="Set1")+
labs(title = "NO emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NO g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2),legend.position = "botto
```



```
ggsave("../../CE594R_data_science_R/figs/NO_conf_comparison.png",width=10,height=8)
```

```
?ggsave
```

```
starting httpd help server ... done
```

How different are the 95% confidence intervals?

How are they different?

Answer: For many of the groups, the changes are either not noticeable. When there are notable differences, the resampling intervals tend to be smaller and include zero less frequently compared to the t-distribution method. Also, the t-distribution bounds are always symmetrical. However, the resampling distributions are not all symmetrical. The upper bounds from the resampling method tend to be at the same location as the t-distribution bounds, but the lower bounds tend to be higher than from the the t-distribution method. This is really evident with the Tim Hwy East Not Compliant confidence bounds.